Tackling the Regulatory Genome

Thesis by

C. Titus Brown

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

2007
(Defended September 5th, 2006)

In memory of Art Kleps (1928 - 1999), Hunter S. Thompson (1937 - 2005), and Hans Bethe (1906 - 2005) – three people who helped me through.

Acknowledgements

Through the years, I have been supported by many great people. Without them none of this work would have happened.

Without my advisor, Eric Davidson, I would not have done most of this work; his intellectual insights and scientific training contributed (and contribute) greatly to my research.

I started doing scientific research in 1993 with Chris Adami, as a Summer Undergraduate Research Fellow. The questions and approaches we discussed during that summer motivated my desire to study biology, and I thank him for his support, his energy, and his ideas.

One of the few good things about spending a long time in graduate school is that you really get to know your thesis committee. In addition to Eric and Chris, Barbara Wold, Paul Sternberg, and Scott Fraser have been wonderful resources and excellent sources of advice. In particular, without my early discussions with Barbara I would not know nearly as much about binding sites and regulatory genomics as I think I do.

I would also like to thank Ellen Rothenberg, Dave McClay, Andy Cameron, Elliott Meyerowitz, and Ray Deshaies for their support, their advice, and their help.

My post-doc advisor Marianne Bronner-Fraser has been supportive beyond reason. Thank you!

I thank Steve Koonin, Curtis Callan Jr., and Jeff Gralnick for their excellent collaborative energies and ideas!

Cathy Yuh taught me how to do molecular biology, and I have had many useful and informative discussions with Paola Oliveri and Andy Ransick about my experiments. Without them, none of my experiments would have worked!

Through the years, Erich Schwarz has been an outstanding dinner companion and an excellent biological check on my more theoretical notions.

The Davidson Lab has been a wonderful place to be, even if I got somewhat grumpy towards the end of my graduate work. I would especially like to thank the unsung heroes of the lab, Jane Rigg and Deanna Thomas. Without Jane and Deanna, there literally would be no lab within which to work!

My family has been the foundation of my life, and without them I would have gone nuts. My parents – old and new – have been great! My sister Claudia has been especially wonderful.

And, last but by no means least, my wife, Tracy Teal, has been a constant source of comfort and support. I cannot imagine living life without her.

Abstract

The structure of the gene regulatory networks that drive animal development is encoded in the genome in *cis*-regulatory regions. Locating these regions and understanding how they integrate regulatory information to produce specific spatiotemporal patterns of gene expression is a major challenge facing developmental biology. This thesis presents computational and experimental work on finding, dissecting, and understanding regulatory regions. I discuss the use of comparative sequence analysis or “phylogenetic footprinting” to locate regulatory regions in animals. I then present experimental work on dissecting the information encoded in the *cyIIIa cis*-regulatory system of the California purple sea urchin, *Strongylocentrotus purpuratus*. Finally, I present a computational investigation of binding site validation techniques in *E. coli*.

Contents

Chapter 1

Introduction

Much of the information in animal genomes necessarily lies in *cis*-regulatory regions (reviewed in (Davidson, 2006)). These regulatory regions contain binding sites that target transcriptional regulators to individual genes, thereby specifying the gene regulatory architecture underlying developmental patterning (Oliveri and Davidson, 2004). However, *cis*-regulatory regions do not simply act as "dumb" integration sites for transcription factors. By targetting the precise genomic binding of spatial activators and repressors, amplitude modulators, and architectural proteins, regulatory regions specify how transcription factors interact to regulate transcription of a specific gene. In effect, regulatory regions can and do act as logic processors, combining their varied transcription factor inputs into novel functional outputs (Yuh et al., 2001, Davidson, 2006).

Uncovering the precise logic function encoded in each *cis*-regulatory region is challenging. Our understanding of transcription factor binding within the genome is still poor, and our understanding of how transcription factors act in combination – the "regulatory code" – is even weaker. Only a few regulatory regions have been seriously investigated in their role as logic processors, and the precise spatiotemporal function encoded by a gene's regulatory DNA has only been largely understood for one gene, *endo16* (Yuh et al., 2001).

In this thesis, I discuss our investigation of the regulatory region of the *cyIIIa* gene in the California purple sea urchin *S. purpuratus*. I also include related work on computational methods for analyzing the "regulatory genome." These works include the use of comparative sequence analysis to discover regulatory regions; a simple technique for recovering conserved regulatory molecules from unassembled genomes and its application to the *S. purpuratus* genome; and a discussion of transcription factor binding site prediction with position-weight matrices in *E. coli*.

1.1 Logic Processing in *cis*-Regulatory regions

Each gene possesses one or more *cis*-regulatory modules. These modules govern where and when the gene is transcribed within the organism. These modules contain many transcription factor binding sites; the factors binding to these sites work together to activate (or repress) the gene in the appropriate place at the appropriate time (reviewed in (Davidson, 2006)).

From the (still relatively few) genes where the global structure of the regulatory modules has been investigated, each regulatory module can be thought of as an independent functional unit. The output of each module, whether it be positive or negative, is combined in a module proximal to the basal promoter of the gene. The output of all of the modules is usually summed, although in some cases the proximal module contains switching logic which alters the logic by which the modules are combined (Yuh et al., 2001).

However, relatively few modules have been analyzed to the degree that we know precisely where the regulatory factors bind, and to what effect.

The *endo16* gene of the *S. purpuratus* is one of the best studied regulatory regions. This regulatory region is contained in a 2.3kb piece of DNA immediately 5' of the transcription start site of *endo16*. The entire region contains over 30 binding sites for 15 transcription factors; these sites have been grouped into 6 modules based on functional analyses. A series of systematic studies of the organization and function of the binding sites has revealed the functions encoded by the two modules closest to the transcription start site, modules A and B. Module A drives initial *endo16* expression in the endoderm, and module B drives late expression more precisely in the midgut of the developing embryo. The complex functional interactions encoded by the binding sites can be represented in a simple logic diagram (see Chapter 6 for a discussion), and the *endo16* gene has served as the central example of the paradigm of *cis*-regulatory information processing.

In Chapter 4, we discuss the regulation of *cyIIIa*, another sea urchin gene expressed in early development. The regulatory region of *cyIIIa* is a 2.3 kb genomic region located immediately upstream of the transcription start site; it contains two modules, A and B, each of which plays a specific and independent regulatory role. Both modules A and B integrate spatial information to produce an aboral ectoderm specific expression pattern; both drive transcription in a distinct temporal pattern.

cis-Regulatory regions perform a important role as information processors, integrating spatial and temporal cues in novel ways to produce specific expression patterns. This role underscores the importance of understanding the precise functions encoded by *cis*-regulatory regions. Ultimately we would like to decode these regions computationally; however, we cannot analyze the functions encoded by *cis*-regulatory regions computationally without first understanding many more *cis*-regulatory regions experimentally (Istrail and Davidson, 2005). The addition of *cyIIIa* to the lexicon of "understood" *cis*-regulatory regions is a significant contribution to this effort.

1.2 Finding Regulatory Regions

Regulatory regions are difficult to find. Unlike protein-coding genes or most non-coding RNA sequences, they have no apparent statistical signature, which renders *ab initio* discovery algorithms ineffective. Regulatory regions tend to be relatively small, between 300 and 1000 bases in length; yet these small regulatory regions can be buried within 20-200 kb of non-coding DNA proximal to each gene. Gene-specific regulatory regions may even reside quite far from the start site of the gene: for example, in the case of mammalian shh, one tissue-specific regulatory module is over 100kb away from the start of transcription (Goode et al., 2005). In smaller genomes such as *C. elegans*, regulatory regions have been found in the introns of neighboring genes. And, in rare cases, binding sites have even been found in protein-coding sequence (Tumpel et al., 2002). All three problems – the large amount of non-coding genomic sequence, the variety of locations in which regulatory regions can reside, and the lack of a clear statistical signature – combine to make locating regulatory regions difficult.

When genomic sequences from two or more closely related species are available, comparative sequence analysis can potentially be used to discover regulatory regions. Comparative sequence analysis relies on detecting the preferential conservation of functionally important DNA in orthologous genomic sequence:

regions of non-coding genomic DNA that are preferentially conserved are good candidates for regulatory elements.

We have successfully used comparative sequence analysis to predict and experimentally verify regulatory regions for a number of sea urchin genes in the Endomesoderm GRN, using Cartwheel and FamilyRelations, a suite of computational tools we developed. The core algorithm used for sequence comparison by Cartwheel is the paircomp algorithm, which does an exhaustive comparison of two sequences with a fixed-width window and a minimum similarity threshold above which to record matches. This comparison, along with any sequence annotations specified by the user, is loaded into the FamilyRelations program. FamilyRelations provides a graphical interface where the user views the analysis results and manipulates analysis parameters.

The first thorough analysis of conserved elements identified by FamilyRelations was done in (Yuh et al., 2002), where 11 of 17 patches conserved between *S. purpuratus* and *L.variegatus* in the *otx* gene region were shown to drive gene expression (Yuh et al., 2002). Since this analysis, many comparative analyses have successfully found regulatory regions in other genes (reviewed in Chapter 2 of this work).

The three software packages that compose our suite of tools are discussed at length in Chapter 3, which was published as a standalone paper in BMC Bioinformatics in 2005 (Brown et al., 2005).

1.3 Finding Regulatory Influences

One of the most pressing issues in building regulatory network models is the problem of finding whcih regulatory molecules bind in regulatory regions of interest. We can *find* regulatory regions relatively easily with comparative sequence analysis, but analyzing the interactions mediated directly by a regulatory region is still quite difficult. Below, I discuss two techniques that bear on the discovery of direct regulatory connections.

1.3.1 Discovering potential upstream regulators using unassembled genomic sequence

Ultimately we would like to understand the role that every regulatory molecule plays in the development of an organism. First, however, we must identify which regulatory molecules are present in that organism. With whole-genome sequencing, this has become feasible.

The suite of transcriptional regulatory factors available in an organism is encoded in the genome. In a newly sequenced genome, transcription factors can usually be identified by sequence homology, because transcription factor binding domains are both ancient and well-conserved. However, most genomes at the moment are neither finished nor particularly well assembled, because whole-genome shotgun sequencing in an affordable coverage range (6-10x) does not usually provide sufficient coverage to assemble the genome unambiguously. Therefore, it is worth considering how to adapt existing computational techniques to work with poorly assembled or even entirely unassembled genomes.

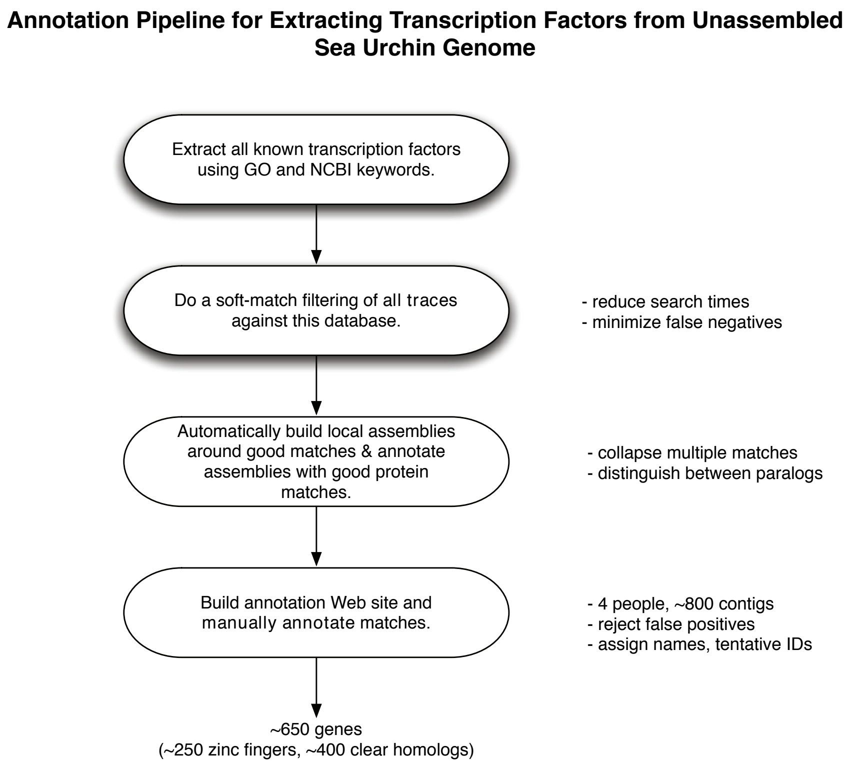

Annotation Pipeline for Extracting Transcription Factors from Unassembled Sea Urchin Genome
Extract all known transcription factors using GO and NCBI keywords.
Do a soft-match filtering of all traces against this database.
- reduce search times
- minimize false negatives
Automatically build local assemblies around good matches & annotate assemblies with good protein matches.
- collapse multiple matches
- distinguish between paralogs
Build annotation Web site and manually annotate matches.
- 4 people, ~800 contigs
- reject false positives
- assign names, tentative IDs
~650 genes
(~250 zinc fingers, ~400 clear homologs)

Figure 1.1: A diagram of our computational pipeline for extracting transcription factors from unassembled genomic sequence.

We developed a simple approach that combined stringent computational searches with manual annotation to produce a high quality list of transcription factors present in the sea urchin genome. The approach involved the following steps when applied to traces from whole-genome shotgun data (diagrammed in figure 1.3.1).

1. Extract all animal transcription factor protein sequences from the NCBI 'non-redundant' database, and combine them with all DNA-binding proteins from the Gene Ontology sequence database;
2. Isolate all genomic traces containing a TBLASTN match to these known transcription factors with an expectation cutoff of $1e-3$.
3. Use BLASTX to retrieve only reciprocal matches. This eliminates some of the false positives characteristic of TBLASTN searches.
4. Group the matches into bins by common matches to 3 independent 20-mer subsequences. Using independent matches makes it possible to bin even somewhat polymorphic traces.
5. Assemble each bin into distinct contigs using the EMBOSS 'merger' program.
6. Annotate assembled contigs by searching them against the NCBI nr database.
7. Use an interactive Web site to manually select contigs of interest, based on their NCBI matches.
8. Design QPCR primers against putative exons.

These primers were then used to determine if the putative exons are expressed early embryogenesis.5

We applied this analysis pipeline to the *S. purpuratus* genome sequence at several stages of assembly. Our initial search for transcription factors was performed against approximately 10 million traces from whole-genome shotgun sequencing (6x coverage), and then cross-checked against several assemblies. Cross-validation with the final published assembly demonstrated that we identified over 95% of the transcription factors identified by the gene prediction pipeline. We also found approximately 15 transcription factors that had been missed by the initial gene prediction pipeline (Sodergren et al., 2006). This success rate suggests that automated assembly of matches around high-complexity protein-coding sequence is a viable technique for initial genome annotation of highly conserved genes.

The results of this computational search, along with further work on expression patterns and some evolutionary comparisons, are detailed in a series of four papers (Howard-Ashby et al., 2006c), (Howard-Ashby et al., 2006b, Howard-Ashby et al., 2006a, Tu et al., 2006). Similar computational approaches were used to identify and characterize genes from other functional categories (Sodergren et al., 2006).

The ultimate use of this search is to determine which regulatory molecules could potentially be involved in gene regulation in early embryogenesis. This considerably reduces the search space for "interesting" regulatory molecules, e.g. those molecules that act as regulatory drivers for genes of interest.

1.3.2 Using position-weight matrices to predict binding sites

The quest to dissect regulatory regions has led to many computational techniques aimed at predicting binding sites. One of the principle search methods used is a weighted matrix representing a "best guess" at the actual binding affinities of a transcription factor. This matrix is known variously as a position-weight matrix (PWM), energy operator, position-specific scoring matrix, or position-specific frequency matrix. In all cases, the matrix represents a scoring function that takes subsequences of a fixed length and assigns them a score; the precise score given, and its numerical meaning, are dependent on the method used to generate the matrix. Most methods of constructing the matrix result in a score representing the similarity of the subsequence to the set of known binding sites.

The theory behind PWMs was first laid out by (Berg and von Hippel, 1987) and (Schneider et al., 1986). The matrix construction method described in (Berg and von Hippel, 1987) uses a logarithmic frequency function that approximates the actual binding energy under two assumptions: first, each nucleotide position in the binding site must contribute *independently* to the interaction; and second, the set of known sites used to construct the matrix must be sampled without bias from the ensemble of possible binding sites.

PWM-based approaches are subject to a number of limitations. When only a few experimentally known binding sites are used to construct a matrix, the resulting matrix tends to have an unacceptably high false positive rate; even when dozens of sites are used, the false positive rate seems to be quite high. Moreover, most known binding sites are aligned around nearly invariant core sequences, but little is known about the precise extent of binding site boundaries; this may result in PWMs that reflect the sequence *necessary* for binding, but not the sequence *sufficient* for binding. Certainly many of the existing PWMs seem to contain insufficient information to precisely specify binding sites on a genomic scale.

Because false positive rates scale with the size of the genome, PWMs have only been generally useful for whole-genome searches in small microbial genomes, although there are notable exceptions to this rule (Mortazavi et al., 2006, Markstein et al., 2002). Even in bacterial genomes, however, false positive rates are high enough that systematic experimental verification of the predictions is generally untenable. As a consequence, two questions arise: is it possible to design a PWM-based procedure that minimizes false positives without reducing the true positive rate? And, if not, is this because the PWM model for locating binding sites is a poor approximation of the true biochemistry of transcription factor binding, or do we simply not have enough biological information to build a precise computational discriminator?

In collaboration with Dr. Curtis G. Callan, I investigated the specificity of PWMs constructed using the known binding sites for the *lacI* and *crp* transcription factors in *E. coli.* Rather than investigating the statistics of the matches themselves relative to a background model, we focused on orthogonal measures of validation, such as distribution of predicted binding sites between coding and non-coding regions and position-specific mutation rates for sites present in both *E. coli* K12 and *S. typhimurium* LT2.

The central discovery of this work (published in (Brown and Callan, 2004), and included in this thesis as Chapter 5) was that many predicted *crp* binding sites are not known but are strongly conserved. We therefore predict that these sites are functional *crp* binding sites, suggesting that there are at least 300 novel binding sites for *crp* in the *E. coli* K12 genome.

This work led to a collaboration with Dr. Jeff Gralnick and Dr. Dianne Newman in which we predicted and then validated a number of *arcA* transcription factor targets in *Shewanella oneidensis* MR1 (Gralnick et al., 2005). A key computational contribution to this work was the discovery that the specificity of the *arcA* weight matrix was strongly enhanced by the inclusion of 10 extra base pairs around the 15 bp core site previously described. We also found that the previously described suite of *arcA* sites in *E. coli* was not in fact specific to *arcA*-regulated genes, and discovered that this earlier work had confused sigma-factor binding sites with the very similar *arcA* binding sites (Brown and Callan, unpublished).

Bibliography

Berg, O. and von Hippel, P., 1987. Selection of dna binding sites by regulatory proteins. statistical-mechanical theory and application to operators and promoters. *J Mol Biol*, **193**(4):723–50.

Brown, C. and Callan, C., 2004. Evolutionary comparisons suggest many novel camp response protein binding sites in escherichia coli. *Proc Natl Acad Sci U S A*, **101**(8):2404–9.

Brown, C., Xie, Y., Davidson, E., and Cameron, R., 2005. Paircomp, familyrelationsii and cartwheel: tools for interspecific sequence comparison. *BMC Bioinformatics*, **6**:70.

Davidson, E., 2006. *The Regulatory Genome*. Academic Press.

Goode, D., Snell, P., Smith, S., Cooke, J., and Elgar, G., 2005. Highly conserved regulatory elements around the shh gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. *Genomics*, **86**(2):172–81.

Gralnick, J., Brown, C., and Newman, D., 2005. Anaerobic regulation by an atypical arc system in shewanella oneidensis. *Mol Microbiol*, **56**(5):1347–57.

Howard-Ashby, M., Materna, S., Brown, C., Chen, L., Cameron, R., and Davidson, E., 2006a. Gene families encoding transcription factors expressed in early development of strongylocentrotus purpuratus. *Dev Biol*, **300**(1):90–107.

Howard-Ashby, M., Materna, S., Brown, C., Chen, L., Cameron, R., and Davidson, E., 2006b. Identification and characterization of homeobox transcription factor genes in strongylocentrotus purpuratus, and their expression in embryonic development. *Dev Biol*, **300**(1):74–89.

Howard-Ashby, M., Materna, S., Brown, C., Tu, Q., Oliveri, P., Cameron, R., and Davidson, E., 2006c. High regulatory gene use in sea urchin embryogenesis: Implications for bilaterian development and evolution. *Dev Biol*, **300**(1):27–34.

Istrail, S. and Davidson, E., 2005. Logic functions of the genomic cis-regulatory code. *Proc Natl Acad Sci U S A*, **102**(14):4954–9.

Markstein, M., Markstein, P., Markstein, V., and Levine, M., 2002. Genome-wide analysis of clustered dorsal binding sites identifies putative target genes in the drosophila embryo. *Proc Natl Acad Sci U S A*, **99**(2):763–8.

Mortazavi, A., Thompson, E., Garcia, S., Myers, R., and Wold, B., 2006. Comparative genomics modeling of the nrsf/rest repressor network: from single conserved sites to genome-wide repertoire. *Genome Res*, **16**(10):1208–21.

Oliveri, P. and Davidson, E., 2004. Gene regulatory network controlling embryonic specification in the sea urchin. *Curr Opin Genet Dev*, **14**(4):351–60.

Schneider, T., Stormo, G., Gold, L., and Ehrenfeucht, A., 1986. Information content of binding sites on nucleotide sequences. *J Mol Biol*, **188**(3):415–31.

Sodergren, E., Weinstock, G., Davidson, E., Cameron, R., Gibbs, R., Angerer, R., Angerer, L., Arnone, M., Burgess, D., Burke, R., *et al.*, 2006. The genome of the sea urchin strongylocentrotus purpuratus. *Science*, **314**(5801):941–52.

Tu, Q., Brown, C., Davidson, E., and Oliveri, P., 2006. Sea urchin forkhead gene family: Phylogeny and embryonic expression. *Dev Biol*, **300**(1):49–62.

Tumpel, S., Maconochie, M., Wiedemann, L., and Krumlauf, R., 2002. Conservation and diversity in the cis-regulatory networks that integrate information controlling expression of hoxa2 in hindbrain and cranial neural crest cells in vertebrates. *Dev Biol*, **246**(1):45–56.

Yuh, C., Bolouri, H., and Davidson, E., 2001. Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development*, **128**(5):617–29.

Yuh, C., Brown, C., Livi, C., Rowen, L., Clarke, P., and Davidson, E., 2002. Patchy interspecific sequence similarities efficiently identify positive cis-regulatory elements in the sea urchin. *Dev Biol*, **246**(1):148–61.

Chapter 2

Finding Regulatory Regions with Comparative Sequence Analysis

2.1 Introduction

Transcriptional regulatory regions are difficult to find. Regulatory regions are relatively short and can lie within tens or hundreds of kilobases of genomic DNA, so experimental techniques to find them are time consuming. As regulatory regions have no clear statistical signature, they cannot be discovered easily with *de novo* computational searches. However, because regulatory regions are the primary drivers of spatiotemporal gene expression during development, techniques that aid in discovering them are important for building models of developmental gene regulatory networks.

One of the most successful computational approaches used to discover regulatory regions is the use of comparative sequence analysis, or "phylogenetic footprinting." Phylogenetic footprinting relies on detecting the signature, or "footprint," of conservation seen in genomic DNA as functional sequence evolves more slowly than non-functional sequence.

In this review, I discuss the use of genomic sequence comparison between closely related species to predict *cis*-regulatory modules. While there have been several reviews of phylogenetic footprinting in vertebrates, there is now a wealth of literature suggesting that phylogenetic footprinting works in most animal model organisms. In light of this literature, I consider several questions: the efficacy of comparative sequence analysis within several phyla; the optimal evolutionary distance for genomic comparisons; and which computational tools are well suited to the task.

First, what do *cis*-regulatory regions look like? We can make several generalizations based on the extensive body of *cis*-regulatory analysis in animals (Wittkopp, 2006, Davidson, 2006). First, each gene may be controlled by one or several *cis*-regulatory modules, although even the simplest regulatory ensembles generally contain two or more modules. Second, most regulatory modules are between 300 bp and 3 kb in size. Third, regulatory modules contain many individual transcription factor binding sites, usually for many different transcription factors. Fourth, regulatory modules for a gene generally reside within the genomic region proximal to that gene. And fifth, regulatory modules tend to work as independent functional modules.

This has several implications for search techniques. We are looking for regions within the genomic neighborhood of the gene under study; thus we can start our search in the 100-200 kb of genomic DNA proximal to the gene (unless the gene is part of a conserved multi-gene complex such as the Hox complex). Because regulatory regions are rich in binding sites, binding site prediction techniques may help guide us to regulatory regions. And, finally, because regulatory modules tend to function independently, we can *test* predicted regions independently, under the assumption that they will work independently.

2.2 Computational techniques for comparing non-coding sequence

There a variety of computational techniques for comparing genomic sequences. The simplest kind of comparison is to compare two genomic sequences. For pairwise comparisons, there are three basic techniques:

all-by-all ungapped comparisons, or dotplots; local alignments; and global alignments. Each of these techniques has advantages and drawbacks; each is capable of detecting or highlighting different sequence features. Below, I discuss these three techniques and computational tools used for each kind of comparison. I also discuss some of the tools and algorithms that can be used to compare three or more sequences.

2.2.1 Dotplots

Dotplots use an ungapped comparison algorithm that compares all subsequences of a given length in one sequence to all subsequences of the same length in another sequence. The similarity of each pair of subsequences is calculated by counting the number of identical bases. Only matches with similarity above a certain threshold are kept and displayed.

Dotplots are the simplest kind of pairwise comparison. The comparison algorithm and scoring system are simple and easy to understand: usually no "score" is calculated beyond the basic similarity measure, and only one parameter – the subsequence length, or windowsize – must be specified by the user. Results can be linked directly and intuitively to the parameters used in the search. Dotplots employ only a single assumption about the evolution of non-coding sequence: sequence similarity implies conservation. Since sequence similarity is calculated based solely on nucleotide identity, with no parameters or feature prioritization, dotplots can detect short ungapped conserved elements that more sophisticated analyses might miss.

However, this simplicity has some drawbacks. All-by-all comparison algorithms tend to be relatively slow compared to the alignment algorithms discussed below, because they are exhaustive. This means that the running times of dotplot algorithms scales with either the product of the sequence lengths or as a power of the window size. In turn, this renders dotplot comparisons slow and memory intensive for large genomic sequences. Alignment algorithms tend to use prefiltering optimizations that make large analyses tractable. Dotplot comparisons also produce increasingly large amounts of data as the windowsizes decrease, because matches are not filtered based on statistical significance or length.

Another significant weakness of dotplot analyses is the difficulty of extending them to multi-way comparisons, although some effort in this direction has been made with the Mussa program.[1]

There are several dotplot comparison programs in regular use for comparative sequence analysis, including DOTTER (Sonnhammer and Durbin, 1995) and seqcomp/paircomp (Brown et al., 2002, Brown et al., 2005).

2.2.2 Local alignment algorithms

Local alignment algorithms build local gapped alignments between two sequences. Generally these alignments are seeded by ungapped identical "words" of a fixed length, and then extended from that point on with a dynamic programming algorithm. Only extensions that increase the score (or keep the score above a minimum) are allowed. This results in an alignment that maps each nucleotide in one subsequence to a

[1] http://woldlab.caltech.edu/cgi-bin/mussa, unpublished.

nucleotide or gap in another subsequence.

Local alignment algorithms are much faster than dotplots because they prefilter sequences based on word identity: for example, because most pairs of short DNA sequences possess no common subsequences of length 11, most of the sequence can be discarded prior to attempting an alignment. This can result in a substantial time savings, at the risk of ignoring small points of similarity between the two sequences.

Like dotplots, local alignments also do an all-by-all comparison of two sequences. This means that a sequence feature on one sequence may match multiple sequence features on another sequence, and local alignment algorithms – like dotplots – can discover inverted or transposed sequence features shared between two sequences. However, as with dotplots, sequence comparisons done with local alignments are difficult to extend to multiple sequences.

One potential disadvantage of local alignments is that results may be sensitive to the parameters used in the search (Dewey et al., 2006). The decision about whether to extend a local alignment across a gap or a region of low sequence similarity depends on the alignment score, which is calculated from a scoring matrix that evaluates similarities and differences as well as a gap opening and extension cost. These parameters need to be adapted to the kind of features being sought; see (Dewey et al., 2006, Wittkopp, 2006, Moses et al., 2006).

BLAST, blastz, BLAT, and SSAHA are four local alignment programs commonly used to compare genomic sequence (Altschul et al., 1997, Schwartz et al., 2003b, Schwartz et al., 2000, Kent, 2002),
(Ning et al., 2001).

2.2.3 Global alignment algorithms

The third technique for comparing two sequences is to build a global alignment between the two sequences. Global alignments provide a one-to-one mapping between each nucleotide on one sequence to either a nucleotide or a gap on the other sequence.

While tools like CLUSTALW and DIALIGN already perform multiple sequence alignments, multiple alignment of large genomic DNA fragments is considerably more computationally challenging than the alignment of relatively short amino acid sequences. A number of tools have been built specifically to address this problem. Most global alignment algorithms are based on constructing an ordered set of local alignments at points of high similarity. These ordered local alignments serve as “anchors” for the alignment of regions of low similarity that lie between the regions of high similarity. For example, AVID uses suffix trees to find the set of all subsequences shared between two sequences, and then selects “anchor” sequences from the set (Bray et al., 2003). The threaded blockset aligner (TBA) and the CHAOS program both build global alignments based on local pairwise alignments (Blanchette et al., 2004, Brudno et al., 2003a). All three programs assume that conserved segments will retain order and orientation between the sequences.

The big advantage of globally aligning two sequences is that detecting conserved regions becomes algorithmically trivial: global alignments provide a linear map of sequence similarity, so detecting regions of high similarity can be done in linear time with respect to the size of the sequences.

Unlike dotplots and local alignments, however, global alignment algorithms cannot usually detect inverted or transposed sequence features. And, as with local alignment algorithms, global alignment algorithms may be quite sensitive to the choice of parameters.

A number of global alignment algorithms are used for genomic-scale comparative sequence analysis. In addition to AVID, TBA, and CHAOS, LAGAN, CLUSTALW, and DIALIGN are commonly used for pairwise global alignments, although CLUSTALW and DIALIGN are too slow to use on large genomic sequences (Brudno et al., 2003b, Thompson et al., 1994, Brudno et al., 2003a).

2.2.4 Comparing more than two sequences

Of the three pairwise comparison techniques discussed above, only global alignment techniques are directly usable for the purpose of comparing more than two sequences. This is because they provide a one-to-one mapping between each pair of sequences, which can easily be extended to additional sequences; with dotplot comparisons and local alignments, it is more difficult to unambiguously connect features across more than two sequences.

There are a number of programs capable of building multiple sequence alignments. Both TBA and CHAOS (discussed above) can be used to align multiple sequences.

Multi-LAGAN uses a progressive alignment technique wherein LAGAN pairwise alignments are iteratively combined into a multiple alignment (Brudno et al., 2003b). As with TBA and CHAOS (above), LAGAN chains local pairwise alignments to speed up the construction of a global pairwise alignment.

MAVID is a multiple sequence alignment tool based on the AVID pairwise alignment tool (Bray and Pachter, 2004). MAVID combines AVID alignments and protein-coding gene predictions to iteratively align sequences along a phylogenetic guide tree, producing progressively more inclusive alignments that parallel the phylogenetic relationships of the sequences to be aligned.

The MultiPipMaker tool processes local alignments created with blastz to iteratively create a global alignment by merging local alignments between a reference sequence and one or more search sequences (Schwartz et al., 2003a). For each search sequence, MultiPipMaker produces a set of non-overlapping local alignments; these several local alignments are then combined into a multiple alignment through another set of iterations that extends multiple alignments locally. The main advantage of this process over many other global alignment algorithms is that only the reference sequence needs to be ordered and oriented, which makes it easy to use MultiPipMaker with draft genome assemblies. (Note that AVID can also order and orient draft sequence, and thus so can MAVID.)

All of these tools approach the problem of multiple sequence alignment primarily from the standpoint of algorithmic efficiency. That is, they attempt to speed up the process of large-scale sequence alignment by anchoring the alignments in local similarities. And despite the variety of approaches, all of these computational tools build a global alignment across multiple sequences, and thus rely on basic co-linearity of conserved features in the sequences being aligned.

2.2.5 Displaying comparison results

There are several ways to display the results of genomic sequence comparisons.

VISTA and MultiPipMaker are two programs that produce plots of sequence similarity relative to a "guide sequence" specified by the user (Mayor et al., 2000, Schwartz et al., 2003a). This guide sequence is usually the reference sequence used for the sequence alignments by AVID (in the case of VISTA) or blastz (for MultiPipMaker). The advantage of this approach is that sequence features can always be linked directly to the coordinates of a single sequence, although in turn the results will be sensitive to the choice of the reference sequence (Chapman et al., 2004, Goode et al., 2005).

SynPlot is another visualization tool for displaying multiple sequence alignments (Chapman et al., 2004). Unlike VISTA and MultiPipMaker, the coordinates used to display the plot are abstract "alignment coordinates" that allow the percentage identities to be calculated across the entire alignment, without requiring a single guide sequence. This allows for straightforward prioritization of peaks of sequence conservation and may be a less biased way of viewing sequence comparison results.

FamilyRelationsII is a stand-alone graphical tool that displays the results of comparisons stored on or performed by the Cartwheel Web server (Brown et al., 2005). Currently it can display pairwise comparisons done with paircomp, blastz, and LAGAN/VISTA, or three-way comparisons done with paircomp. One singular advantage of FamilyRelationsII is that it lets users quickly and easily compare analyses performed with a dotplot analysis (paircomp), a local alignment program (blastz), and a global alignment program (VISTA).

Most genome viewers (e.g. both the UCSC genome site and ENSEMBL) include representations of conservation or synteny in their static views, as well (Hinrichs et al., 2006, Hubbard et al., 2006). These analyses are usually based on whole genome comparisons done with blastz.

2.3 Comparative sequence analysis in vertebrates

The vast majority of efforts to use comparative sequence analysis to find regulatory regions have taken place in the vertebrates. Vertebrate genomes are much larger than the worm and fly genomes, making it much more difficult to find regulatory regions experimentally. Thus computational aids are more necessary. Moreover, the medical and agricultural relevance of many vertebrates means that many vertebrate genomes sequences are available, so there is an opportunity to look at conservation across many evolutionary distances. Below I describe a number of vertebrate analyses, chosen based on the amount of experimental validation done and on the number of species analyzed.

The SCL locus The stem cell leukemia gene (SCL) encodes a transcription factor involved in regulating hemopoiesis and vasculogenesis (Barton et al., 2001). Both its embryonic and adult patterns of expression are highly conserved across the vertebrates, including mammals and the teleosts (Barton et al., 2001).

The SCL locus is an ideal genomic region for studying the effectiveness of comparative sequence analysis across the mammals and the vertebrates. Many of the murine regulatory regions have already been identified, and the expression pattern is conserved across the vertebrates, suggesting that the SCL regulatory sequences may also be conserved. Moreover, the genomic locus is relatively simple: there are no SCL homologs located near to the SCL gene in mouse, and there is no synteny between the mouse SCL locus and the pufferfish SCL locus, implying that proximal enhancers may be responsible for the conserved SCL expression pattern (Barton et al., 2001, Gottgens et al., 2001). Finally, several regions sensitive to either DNAseI or restriction endonuclease digestion have been identified within the 40kb surrounding the SCL gene, thus establishing a number of potential regulatory regions experimentally (Gottgens et al., 2001).

A comparison of the SCL locus across four mammals (mouse, human, dog, and rat) and four other vertebrates (chick, *Fugu*, *Tetraodon*, and zebrafish) showed that all eight known mouse regulatory regions were conserved between the mammals, while five of the eight regulatory regions were conserved between the mammals and chicken (Chapman et al., 2004, Gottgens et al., 2001). Only two of the eight mouse regulatory regions were found in *Fugu*, and none at all were identifiable in zebrafish. This is especially surprising because a 10.4kb construct containing only the *Fugu* SCL genomic sequence and its surrounding intergenic DNA is capable of driving correct expression in zebrafish (Barton et al., 2001). Thus while regulatory *function* of the SCL genomic sequence is similar in *Fugu*, and the upstream regulatory factors must be present in zebrafish, the regulatory sequence has diverged. A strong conclusion from this work is that the proper choice of organisms for phylogenetic footprinting relies on more than conservation of expression patterns.

(Chapman et al., 2004) also used multiple sequence alignments among four mammals (mouse, human, dog, and rat) to assess the conservation of transcription factor binding sites in the SCL locus. All 19 experimentally known binding sites are in sequence elements that are conserved between all four mammal sequences, and the specificity of detection improves dramatically as sequence is eliminated by including more species. Moreover, 17 of the 19 known binding sites occurred in exactly conserved blocks (no gaps or mismatches) of seven base pairs. This suggests that aligning multiple closely related sequences, as with the mammalian sequences, may be an effective general technique to constrain the search space for binding sites.

Interleukins at 5q31 The regulation of three biomedically important cytokines, *interleukin-4, interleukin-13*, and *interleukin-5* was investigated using mouse/human comparisons (Loots et al., 2000). Earlier work with transgenic yeast artificial chromosomes established that the regulatory elements responsible for type 2 T helper cell expression were contained within the same 1mb 5q31 locus as the three genes (Symula et al., 1999). Thus upstream regulatory linkages are conserved between mouse and human, suggesting that regulatory sequence may also be conserved.

A VISTA-style scan for 100bp elements conserved between mouse and human at 70% sequence similarity identified 245 conserved elements, of which 145 overlapped known coding sequences (Loots et al., 2000). Of the remaining 90 conserved non-coding elements, one was an already known regulatory module controlling the co-regulation of *GM-CSF* and *IL-3*. Moreover, the presence of 15 elements was examined in humans and other mammals. 12 of the 15 elements were present in single copies in human, and 10 elements were present in the genomes of at least two other mammals, although their genomic locations in these other mammals were not identified and so direct conservation was not confirmed.

The CNS-1 conserved non-coding region, a 401 bp sequence located in the intergenic region between *IL-4* and *IL-13*, was investigated further. When a 450kb YAC transgene lacking this region was introduced into mice and purified naive T cells from these mice were induced *in vitro*, significantly fewer IL-4 and IL-13 producing cells were produced than when cells containing the wild type YAC transgene were induced. However, actual expression levels per cell did not vary from the wild type, suggesting that the CNS-1 region regulates the activation of these genes on a cell-by-cell basis, perhaps through modifications to the large-scale chromatin structure. The expression level of IL-5 also decreased in cells without CNS-1, while two other genes in the same region expressed at the same level independently of its presence. Thus it seems that CNS-1 is specific to interleukin gene regulation.

α-globin The human α-globin gene cluster is another well-studied region that has been used to assess the effectiveness of comparative sequence analysis for the discovery of regulatory regions (Flint et al., 2001, Hughes et al., 2005). As of 2005, the α-globin region had been identified and sequenced in 22 vertebrates (18 mammals, 3 fishes, and chick). (Hughes et al., 2005) compared conservation patterns in these 22 regions with previously known data on DNAseI hypersensitivity, nuclear localization, nuclear matrix attachment patterns, DNA replication patterns, chromatin structure and modification, and gene regulatory elements.

Candidates for regulatory elements in the α-globin region were produced by inspecting blastz local alignments generated by PipMaker (Hughes et al., 2005). After known coding sequences were removed, only sequences present in more than 75% of mammalian species were kept. These regions were then extracted and aligned using CLUSTALW, DIALIGN, and BESTFIT. Only those sequences containing at least one 10-base window where 6 of the 10 bases were identical across 75% of the mammalian species were kept as putative “multi-species conserved sequences,” or MCSs. Using these criteria, a total of 24 MCSs were identified in the 238 kb surrounding the human α-globin genes.

The correlation between conservation and function was excellent. 15 of the 24 MCSs were already known regulatory regions, and two splicing regulatory sequences and 4 previously unknown exons were also identified. Only three of the 24 MCSs identified in this study could not be assigned a probable function. Overall, all but one already known regulatory region was identified with this approach, suggesting that comparative sequence analysis can be both very sensitive and quite specific.

Two other programs, WEBMCS and GUMBY, were also used to analyze this region (Margulies et al., 2003, Prabhakar et al., 2006). WEBMCS successfully identified 12 of the 24 MCSs as well as one additional element that failed the MCS criteria above; GUMBY identified 20 of the 24 MCSs and 19 additional candidates.

sox2 The chicken *sox2* gene is expressed widely in the nervous system and is implicated in neural development. (Uchikawa et al., 2003) used a reporter system in chick to systematically scan a 50 kb region of the *sox2* locus for positive regulatory activity. After isolating 10 distinct regions that drove expression in neural tissues, they used MacVector and VISTA to compare the isolated regions with sequence from the orthologous human and mouse *sox2* regions. All of the experimentally verified regulatory regions were indeed conserved between human, mouse, and chick, although there were many additional conserved elements that had not been tested.

A similar approach was used to identify four tissue-specific regulatory sequences in the *N-cadherin* chick locus (Matsumata et al., 2005). In this study, (Matsumata et al., 2005) generated a random library containing subfragments of the 219 kb *N-cadherin* genomic region inserted into a reporter construct. Individual clones from this library were tested for function in cell lines and chick embryos, resulting in five functional regulatory regions. Of these five regions, four directed tissue-specific and one enhanced expression in a spatially neutral manner. The positive regions were sequenced and compared with the human *N-cadherin* region for conservation patterns using VISTA. Of the four tissue-specific regulatory regions discovered experimentally, three contained patches strongly conserved between human and chick.

gdf6 GDF6 is a member of the bone morphogenetic protein (BMP) signalling family. It plays a role in many diverse vertebrate developmental processes, including the formation of skeletal joints, limb, ear, and skull joints (Settle et al., 2003). A series of studies located a number of distinct regulatory modules that drive expression in subsets of the tissues in which *gdf6* is normally expressed (Mortlock et al., 2003, Mortlock et al., 2004, Portnoy et al., 2005). These regulatory modules are spread across more than 250 kb of genomic DNA in the mouse, and are individually conserved within different subsets of the gnathostomes, amniotes, mammals, and eutherian mammals (Portnoy et al., 2005).

As with *sox2* and *N-cadherin* (above), many of the enhancers were initially identified experimentally, and comparative sequence analysis between mouse and human genomic sequence was then used to pinpoint the precise location of each enhancer. An initial study used overlapping series of mouse BAC-lacZ knockin constructs to identify the broad regions within which certain tissue-specific enhancers resided (Mortlock et al., 2003). The sequence of the entire mouse BAC was then determined and compared with the human sequence of the *gdf6* locus. Several different comparative techniques were used to look for conserved sequence elements, including VISTA, PipMaker, and the UCSC genome browser's annotated L-score alignments (Mayor et al., 2000, Schwartz et al., 2000, Waterston et al., 2002). These techniques all returned very similar results. A number of highly conserved elements found by VISTA were isolated and tested for function, initially by deletion and then later with reporter constructs (Mortlock et al., 2003, Portnoy et al., 2005).

Many of the elements were found to be both necessary and sufficient for gene expression in a variety of tissues, including the mammary gland, genital tubercle, larynx, and dorsal neural tube.

Another study used sequence from 13 additional species spanning the gnathostomes – zebrafish, *Xenopus tropicalis*, *Fugu*, chick, platypus, and 7 additional eutherian mammals – to investigate patterns of conservation at the *gdf6* locus both across the vertebrates and within vertebrate subphyla (Portnoy et al., 2005). In addition to a MultiPipMaker analysis, several other tools were used to identify conserved sequences across these alignments, including ExactPlus and WebMCS (Antonellis et al., unpublished; (Margulies et al., 2003)). While several of the known functional regulatory elements were conserved between human and fish, others displayed more restricted patterns of conservation limited to the amniotes or the eutherian mammals (Portnoy et al., 2005). Almost all of the conserved sequences found in this study were detectable with fairly stringent search parameters.

Hox clusters The Hox clusters contain large syntenic blocks of genes conserved across the vertebrates, and one of the reasons behind this syntenic association may be that many interdigitated regulatory elements are present (Lemons and McGinnis, 2006). Consequently many comparative sequence analysis studies have been used to discovered regulatory elements in the Hox regions (reviewed in (Lee et al., 2006)). In particular, (Lee et al., 2006) do an exhaustive comparison of the *Fugu* Hox regions with the homologous human and mouse Hox regions with VISTA and show that non-coding sequences conserved between *Fugu* and mammals contain many of the previously identified regulatory elements.

apo(a) In a study of the regulatory region of the apolipoprotein (a) gene, (Boffelli et al., 2003) used a novel approach that they termed "phylogenetic shadowing". Because the apo(a) gene is confined to primates, the standard comparative sequence analysis approaches discussed above that rely on divergence of nonfunctional sequence would not have identified functionally significant sequence. Instead, (Boffelli et al., 2003) analyzed a multiple sequence alignment of highly conserved sequences taken from a variety of primates including Old World monkeys and hominoids. Based on this alignment, they classified sequences as "fast evolving" or "slow evolving", and examined only those that were classified as slow evolving, i.e. those that were predicted to be under selective constraint. These sequences were experimentally tested and found to be positive regulatory elements containing binding sites for factors present in nuclear extracts from a liver cell line.

One intriguing result from this work was that phylogenetic shadowing cleanly delineated relatively short subsequences under selection. Phylogenetic shadowing may thus be more capable of distinguishing individual binding sites or functional subunits within modules than other comparative sequence analysis techniques.

Systematic analysis of human-*Fugu* or human-mouse-rat conserved sequences While most analyses focus on finding regulatory regions surrounding single genes or contained in small gene regions, (Pennacchio et al., 2006) did a systematic assay of 167 human sequences that were conserved either between human, mouse, and rat, or between human and *Fugu*. Sequence elements were chosen based on a VISTA-like analysis with a minimum of 70% conservation between either human and *Fugu* or human, mouse, and rat.

Of 83 sequences conserved strongly between human and *Fugu*, 24 (29%) functioned as positive tissue-specific regulatory modules in E11.5 stage mice. The same assay found that 51 of 84 (61%) of elements that were "ultraconserved" between human, mouse and rat were positive regulatory elements. 54 of the human/mouse/rat ultraconserved sequences were also present in *Fugu*, and 33 of these 54 sequences (61%) drove expression. Overall, 40% of sequences conserved between human and *Fugu* were functional (57 of 137), while 61% of sequences conserved between human, mouse and rat were functional (51 of 84).

These results suggest that conservation among multiple closely-related species may be a better positive predictor for regulatory elements than pairwise comparisons between very distant species.

2.4 Invertebrate sequence comparisons

Many more vertebrate genomes have been sequenced than invertebrate genomes, and the invertebrate literature on comparative sequence analysis is consequently much smaller. Most work on the discovery of *cis*-regulatory regions with comparative sequence analysis in invertebrates has been done in sea urchins, ascidians, and nematodes.

2.4.1 Comparative sequence analysis effectively identifies regulatory regions in ascidians

Comparative sequence analysis between the two fully sequenced ascidians, *Ciona intestinalis* and *Ciona savignyi*, has been effective at helping to analyze *cis*-regulatory regions. Because these genomes are relatively compact, finding *cis*-regulatory regions has been easy without comparative sequence analysis (see (Kusakabe, 2005) for a review). Comparative sequence analysis in ascidians has primarily been used to narrow the search for binding sites.

(Yagi et al., 2004, Awazu et al., 2004, Bertrand et al., 2003) all used known functional regions from *C. intestinalis* as a basis for comparison with *C. savignyi*. In all three cases, comparative sequence analysis identified strongly conserved sub-elements of the functional region that acted as independent enhancers. In each case, when these regions were examined for binding sites, known or suspected binding sites were identified on the basis of conservation. These binding sites were then shown to be functional.

(Christiaen et al., 2005) used a VISTA analysis to identify multiple conserved non-coding elements in the 20kb *pitx* locus. Based on a minimum of 65% or higher identity across 80 bp, (Christiaen et al., 2005) identified 10 putative *cis*-regulatory modules spread throughout the *pitx* upstream region and introns. They tested five of these elements for function and showed that three distinct elements cooperated to drive expression of the *pitx a/b* isoform in the anterior neural boundary and stomadaeum.

A larger study by (Johnson et al., 2004) showed that conservation of non-coding sequence between *C. intestinalis* and *C. savignyi* correlated well with regulatory function for several different genes. Regulatory elements for six distinct *C. intestinalis* genes – *α-tubulin*, *noto9*, *forkhead*,*brachyury*, *troponin I*, and *synaptotagmin* – were tested in *C. savignyi* and shown to be functional, which correlated well with 5' sequence conservation for these genes (Johnson et al., 2004).

While *C. intestinalis* and *C. savignyi* are sufficiently close that comparative sequence analysis can be used to find regulatory elements, regulatory elements in a third ascidian model organism, *Halocynthia roretzi*, may be too divergent to recognize by direct sequence comparison. Comparison of *brachyury cis*-regulatory sequence and expression in *C. intestinalis* and *H. roretzi* suggests that while the role of *brachyury* in the presumptive notochord is conserved, neither the *cis*-regulatory sequence responsible for the expression nor the direct upstream connections are conserved (Takahashi et al., 1999). Likewise, a study of regulation of the *otx* gene found that despite possessing several direct regulators in common, the *H. roretzi* and *C. intestinalis otx cis*-regulatory regions shared no clear sequence similarity (Oda-Ishii et al., 2005). These studies suggests that *C. intestinalis* and *C. savignyi* are currently the best ascidian genome pair to use for comparative sequence analysis.

2.4.2 Comparative sequence analysis effectively identifies regulatory regions in sea urchins

A number of regulatory elements have been discovered by finding regions conserved between *S. purpuratus* and *L. variegatus*, two species of sea urchins approximately 50 my diverged. The first analysis between these species was done on the otx gene using the seqcomp and FamilyRelations software (Brown et al., 2002, Yuh et al., 2002, Brown et al., 2005). This software suite does a dotplot-style exhaustive comparison of two regions, using a fixed-width window of a size specified by the user. Only those windows with the number of matches above a user-defined threshold are reported and displayed.

In the *otx* analysis, a 70%/50 bp window analysis located 17 patches of sequence similarity indicative of conservation. These patches were then tested for positive regulatory function. Of these 17 patches, 11 drove transcription (Yuh et al., 2002). Since the analysis of *otx*, several additional studies have been published, including an analysis of the regulation of *wnt8* (Minokawa et al., 2005); *delta* (Revilla i Domingo et al., 2004); *blimp1/krox* (Livi and Davidson, 2007); and *gcm* (Ransick and Davidson, 2006).

A striking feature of all of these analyses was the strong conservation of the regulatory modules: all of the regulatory modules isolated by comparative analyses were conserved at over 90% across several hundred bases, with very few insertions or deletions (indels). This suggested that indels either were not a prevalent mode of sequence mutation, i.e. that they either did not occur, or were strongly selected against when they did occur. To investigate this, (Cameron et al., 2005) examined sequence surrounding known regulatory elements from *S. purpuratus*, *L. variegatus*, and *S. franciscanus*, a sea urchin more closely related to *S. purpuratus* than to *L. variegatus*. They found that indels in nonfunctional sequence occurred frequently, while indels were effectively absent in the known regulatory sequences. This suggests that regulatory modules are refractory to indels.

Despite common trans-regulatory connections, comparisons between two distant echinoderms, *S. purpuratus* and *A. miniatia*, failed to detect conservation in the *cis*-regulatory sequence (V. Hinman, pers. communication).

2.4.3 Comparative sequence analysis in *C. elegans*

There are relatively few published cases of regulatory region discovery using simple sequence identity between the two sequenced nematode genomes, *C. elegans* and *C. briggsae*. Many of the cases involve dot-plot style comparisons done with Dotter or FamilyRelationsII (Thacker et al., 1999, Kirouac and Sternberg, 2003, Teng et al., 2004). In only one of the analyses, that of *egl-5*, an *Abd-B* homolog, were large regions of conservation detected (Thacker et al., 1999). In both of the other analyses, (Thacker et al., 1999) and (Kirouac and Sternberg, 2003), a windowed comparison was used to locate 50-100 bp fragments of sequence with strong similarity between *C. elegans* and *C. briggsae*. In all three cases, the conserved elements were shown to be functional.

2.4.4 The strange case of *Drosophila melanogaster*

There are relatively few works describing the use of sequence comparison alone to find regulatory regions *de novo* in *D. melanogaster* (reviewed in (Wittkopp, 2006)). In fact, several different studies have assessed the conservation of known *Drosophila melanogaster* regulatory regions in *D. pseudoobscura* and concluded that regulatory sequence is not strongly conserved over nonfunctional sequence in many cases (Berman et al., 2004, Emberly et al., 2003, Wittkopp, 2006). However, preservation of binding site presence and/or predicted binding near to genes turns out to be a powerful predictor of regulatory function (Sinha et al., 2004, Macdonald and Long, 2005, Grad et al., 2004). This is very different from the examples discussed above for vertebrate, sea urchin, and ascidian comparisons, where sequence identity in noncoding regions is an excellent indicator of regulatory function.

One possible reason for poor conservation of *cis*-regulatory regions is that existing alignment tools do a poor job of building alignments in *Drosophila* sequences. Recent work on parametric alignment techniques has demonstrated that when using optimal Needleman-Wunsch alignments, known binding sites are more conserved than previously thought (Dewey et al., 2006). However, there are many observations of rapid sequence evolution in regulatory regions suggesting that in fact regulatory regions are evolving fast (reviewed in (Wittkopp, 2006)).

Many of the known enhancers used to benchmark conservation in *Drosophila* appear to be undergoing fairly rapid sequence evolution (Wittkopp, 2006, Moses et al., 2006). Some studies had already reached this conclusion by doing directed sequencing of orthologs of well-known gene regions. In particular, the *eve* stripe 2 study by (Ludwig et al., 1998) concluded that the *eve* stripe 2 *cis*-regulatory sequence had diverged between *D. pseudoobscura* and *D. melanogaster*, even though the function of the enhancer remained similar.

2.5 Discussion

2.5.1 What genomes should be used for comparison purposes?

The choice of genomes when using comparative sequence analysis to search for regulatory regions is important. For some model organisms, there is only one species pair, e.g. *C intestinalis* and *C. savignyi* are the only two fully-sequenced ascidian genomes available, and likewise *C. elegans* and *C. briggsae* are the only two nematode genomes currently available.

In some cases, only one genome is fully sequenced and genomic sequence for a nearby species must be obtained from a large insert library. This approach has been particularly successful for *S. purpuratus* and *L. variegatus*, two sea urchins.

There are now many vertebrate genomes available, and it can be difficult to choose which ones to use for comparative sequence analysis. Results from two papers – the Pennacchio et al. investigation of genome-wide conservation, and the phylogenetic shadowing analysis of the apo(a) gene – suggest that comparisons between sequences from multiple genomes can be an effective technique. To locate mouse regulatory regions, comparisons between human, mouse and rat are particularly effective. The *sox2* and *N-cadherin* work by Kondoh et al. suggests that mouse or human genomic sequences are effective partners when looking for neural regulatory elements in chick. This is likely to hold true for *Xenopus*, too, which is not substantially more distant from human than is chick.

Despite the successes in using *Fugu* or zebrafish as partners to find regulatory elements in mouse and human (see (Venkatesh and Yap, 2005) for a review), the systematic investigation by Pennacchio et al. demonstrates that *Fugu* is a less effective partner for mouse than are multiple mammalian sequences. There are two possible reasons for this: first, fish are quite diverged – $\tilde{4}$50 my – from mammals, almost double the divergence times of avians. This means that regulatory elements may have diverged at the sequence level, even if the expression pattern and upstream regulators are conserved (see the case of SCL, above). Second, the teleosts (including *Fugu* and zebrafish) have undergone an additional round of genome duplication, which means that the regulation of paralogs may have diverged, complicating the search for conserved regulatory elements. Overall, *Fugu* and zebrafish are probably not as useful for comparative sequence analysis as are multiple mammal sequences and chick or *Xenopus*.

Drosophila is a case where direct comparative sequence analysis frequently does not help identify regulatory elements, so even though there are now many *Drosophila* genomes available, it is unclear which (if any) of them are most effective as partners.

2.5.2 How many genome sequences should be compared?

There are conflicting reports on the utility of multiple sequence comparison. In vertebrate genomes, (Margulies et al., 2006) show that aligning multiple sequences with the threaded blockset aligner can add significantly to the amount of alignable sequence, because similarity shared between only two sequences is progressively included in the multiple sequence alignment. However, (Moses et al., 2006) test multiple alignment algorithms against a simulation of molecular evolution and show that the efficacy of multiple alignment algorithms is determined almost entirely by the evolutionary distance of the maximally divergent pair of sequences. This suggests that multiple sequence alignment algorithms do not effectively incorporate information from multiple sequences. The reasons for this disparity are not clear; a major difference between the studies is that the (Margulies et al., 2006) study is based in actual vertebrate sequence, while (Moses et al., 2006) work with parameters taken from *Drosophila* sequence.

Nonetheless, it is clear from empirical data – the apo(a) and Pennacchio et al. study, in particular – that comparing multiple sequences can aid in discovering functional regulatory elements in vertebrates.

2.5.3 What tools are best to use?

One important question to ask is how well the various tools perform on actual genomic sequence. Unlike the better studied question of protein sequence alignment, there are relatively few ways to calculate a "gold standard" against which to compare the output of alignment programs. For example, (Pollard et al., 2004) benchmark a variety of alignment programs using simulated sequence evolution and determine that both local and global alignment algorithms are effective at detecting evolutionarily constrained sequence (also see the correction to (Pollard et al., 2004), (Pollard et al., 2006)), suggesting that most alignment programs have acceptable sensitivity. However, it is unclear how specific these programs are in finding regulatory elements. So far, no large scale comparison has been done against a set of known regulatory elements. So, the best way of measuring the effectiveness of sequence comparison algorithms is probably to look anecdotally at the many examples of regulatory elements that have been discovered.

Perhaps the most commonly used comparative sequence analysis programs is VISTA. There are several *a priori* reasons why VISTA might not be the best program to use to discover regulatory elements: the VISTA algorithm relies on large windowsizes (75-100 bp) and requires highly conserved features that are co-linear. However, in practice VISTA and other global alignment algorithms have been very effective in finding functional regulatory elements, e.g. as in the gdf6 and Pennacchio studies above. Global alignment algorithms work well because *cis*-regulatory elements do tend to be highly conserved across several hundred bases and also tend to be conserved co-linearly. The only drawback to using VISTA is that VISTA requires that the user pick a reference sequence, and the quality of the resulting alignment may be sensitive to this choice (Chapman et al., 2004, Goode et al., 2005).

When draft quality sequence is the only sequence available, VISTA may not be as effective as other programs, because the global alignment algorithm that underlies VISTA only works when the sequences are ordered and oriented. In this case, an all-by-all analysis with PipMaker/blastz or FamilyRelations/seqcomp will be more capable of detecting features that are out of order or inverted.

The question of whether to use multiple alignment software depend critically on the availability of more than two relevant genomic sequences. Pairwise comparisons in ascidians, sea urchins, and nematodes have been very effective in detecting regulatory elements. However, work in mammals suggests that comparing conservation among three or more sequences may help prioritize conserved elements for experimental assays and can also help detect binding sites. In situations where more than two sequences are available, VISTA, MultiPipMaker, and SynPlot are all effective "off the shelf" choices, although more specialized approaches such as that used in the apo(a) study may also be useful.

2.5.4 How should the genomic regions to be compared be chosen?

Deciding what region to examine for the presence of regulatory regions relevant to the gene of interest can be difficult: there are many situations in which regulatory regions are quite distal to the gene they regulate. Synteny, the retention of gene presence and order on a large scale, can provide clues as to the scale on which a search should be conducted, because regulatory regions usually need to remain relatively close to the gene they regulate. If a regulatory region for one gene lies proximal to or within a neighboring gene, then there is a selective pressure to retain the relationship between the two genes (Mackenzie et al., 2004).

Several studies have shown a correlation between synteny and regulatory sequences. For example, (Goode et al., 2005) showed that there were a number of conserved *cis*-regulatory elements in the 4 megabase shh-containing 7q36.3 region in human, which may be responsible for the conserved gene structure of this region between human, mouse, and fugu. Likewise, (Poulin et al., 2005) found several regulatory regions for the Dachshund gene spread across the surrounding 2mb, which contained no other genes. This suggests that regulatory regions are responsible for the conservation of structure in the Dachshund locus.

(Ahituv et al., 2005) showed that, on average, the length of human-mouse-chick or human-mouse-frog syntenic blocks correlated with the number of conserved non-coding elements present in those blocks, while the number of genes present in syntenic blocks decreased. They conclude that this correlation suggests the presence over conserved long-range regulatory interactions. Therefore syntenic blocks may be good "units of comparison" when searching for regulatory regions.

Even when direct sequence comparison detects little or no conservation, synteny may be useful. In the case of SCL (above) the lack of multi-gene synteny between mouse and fugu indicated that were regulatory regions conserved, they would almost certainly be in the intergenic or intronic sequence around the SCL gene. This is the case, although ironically the sequence of most of the enhancers in the 10.4kb fugu genomic region are not actually conserved with mouse – despite directing correct expression of a transgene in zebrafish. Thus while fugu sequence may sometimes be too distant for direct discovery of noncoding sequences conserved with mammals, the presence or lack of synteny between mammals and fugu can be informative.

2.6 Conclusions

Comparative sequence analysis is an effective search technique for regulatory elements in model organisms when one or more genomic sequences from nearby species are available. Comparative sequence analysis can be a sensitive and specific way to search for regulatory elements: features conserved between closely related species correlate well with known functional elements, and at least in human and mouse, over 60% of sequences conserved between human, mouse and rat function as tissue-specific regulatory elements. Most of the programs available for comparative sequence analysis perform well, and VISTA is an easy and accurate first choice.

Bibliography

Ahituv, N., Prabhakar, S., Poulin, F., Rubin, E., and Couronne, O., 2005. Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny. *Hum Mol Genet*, **14**(20):3057–63.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D., 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, **25**(17):3389–402.

Awazu, S., Sasaki, A., Matsuoka, T., Satoh, N., and Sasakura, Y., 2004. An enhancer trap in the ascidian ciona intestinalis identifies enhancers of its musashi orthologous gene. *Dev Biol*, **275**(2):459–72.

Barton, L., Gottgens, B., Gering, M., Gilbert, J., Grafham, D., Rogers, J., Bentley, D., Patient, R., and Green, A., 2001. Regulation of the stem cell leukemia (scl) gene: a tale of two fishes. *Proc Natl Acad Sci U S A*, **98**(12):6747–52.

Berman, B., Pfeiffer, B., Laverty, T., Salzberg, S., Rubin, G., Eisen, M., and Celniker, S., 2004. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in drosophila melanogaster and drosophila pseudoobscura. *Genome Biol*, **5**(9):R61.

Bertrand, V., Hudson, C., Caillol, D., Popovici, C., and Lemaire, P., 2003. Neural tissue in ascidian embryos is induced by fgf9/16/20, acting via a combination of maternal gata and ets transcription factors. *Cell*, **115**(5):615–27.

Blanchette, M., Kent, W., Riemer, C., Elnitski, L., Smit, A., Roskin, K., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E., *et al.*, 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, **14**(4):708–15.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K., Ovcharenko, I., Pachter, L., and Rubin, E., 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**(5611):1391–4.

Bray, N., Dubchak, I., and Pachter, L., 2003. Avid: A global alignment program. *Genome Res*, **13**(1):97–102.

Bray, N. and Pachter, L., 2004. Mavid: constrained ancestral alignment of multiple sequences. *Genome Res*, **14**(4):693–9.

Brown, C., Rust, A., Clarke, P., Pan, Z., Schilstra, M., Buysscher, T. D., Griffin, G., Wold, B., Cameron, R., Davidson, E., *et al.*, 2002. New computational approaches for analysis of cis-regulatory networks. *Dev Biol*, **246**(1):86–102.

Brown, C., Xie, Y., Davidson, E., and Cameron, R., 2005. Paircomp, familyrelationsii and cartwheel: tools for interspecific sequence comparison. *BMC Bioinformatics*, **6**:70.

Brudno, M., Chapman, M., Gottgens, B., Batzoglou, S., and Morgenstern, B., 2003a. Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics*, **4**:66.

Brudno, M., Do, C., Cooper, G., Kim, M., Davydov, E., Green, E., Sidow, A., and Batzoglou, S., 2003b. Lagan and multi-lagan: efficient tools for large-scale multiple alignment of genomic dna. *Genome Res*, **13**(4):721–31.

Cameron, R., Chow, S., Berney, K., Chiu, T., Yuan, Q., Kramer, A., Helguero, A., Ransick, A., Yun, M., and Davidson, E., *et al.*, 2005. An evolutionary constraint: strongly disfavored class of change in dna sequence during divergence of cis-regulatory modules. *Proc Natl Acad Sci U S A*, **102**(33):11769–74.

Chapman, M., Donaldson, I., Gilbert, J., Grafham, D., Rogers, J., Green, A., and Gottgens, B., 2004. Analysis of multiple genomic sequence alignments: a web resource, online tools, and lessons learned from analysis of mammalian scl loci. *Genome Res*, **14**(2):313–8.

Christiaen, L., Bourrat, F., and Joly, J., 2005. A modular cis-regulatory system controls isoform-specific pitx expression in ascidian stomodaeum. *Dev Biol*, **277**(2):557–66.

Davidson, E., 2006. *The Regulatory Genome*. Academic Press.

Dewey, C., Huggins, P., Woods, K., Sturmfels, B., and Pachter, L., 2006. Parametric alignment of drosophila genomes. *PLoS Comput Biol*, **2**(6):e73.

Emberly, E., Rajewsky, N., and Siggia, E., 2003. Conservation of regulatory elements between two species of drosophila. *BMC Bioinformatics*, **4**:57.

Flint, J., Tufarelli, C., Peden, J., Clark, K., Daniels, R., Hardison, R., Miller, W., Philipsen, S., Tan-Un, K., McMorrow, T., *et al.*, 2001. Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. *Hum Mol Genet*, **10**(4):371–82.

Goode, D., Snell, P., Smith, S., Cooke, J., and Elgar, G., 2005. Highly conserved regulatory elements around the shh gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. *Genomics*, **86**(2):172–81.

Gottgens, B., Gilbert, J., Barton, L., Grafham, D., Rogers, J., Bentley, D., and Green, A., 2001. Long-range comparison of human and mouse scl loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. *Genome Res*, **11**(1):87–97.

Grad, Y., Roth, F., Halfon, M., and Church, G., 2004. Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in drosophila melanogaster and d.pseudoobscura. *Bioinformatics*, **20**(16):2738–50.

Hinrichs, A., Karolchik, D., Baertsch, R., Barber, G., Bejerano, G., Clawson, H., Diekhans, M., Furey, T., Harte, R., Hsu, F., *et al.*, 2006. The ucsc genome browser database: update 2006. *Nucleic Acids Res*, **34**(Database issue):D590–8.

Hubbard, T., Aken, B., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., *et al.*, 2006. Ensembl 2007. *Nucleic Acids Res*, .

Hughes, J., Cheng, J., Ventress, N., Prabhakar, S., Clark, K., Anguita, E., Gobbi, M. D., de Jong, P., Rubin, E., and Higgs, D., *et al.*, 2005. Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proc Natl Acad Sci U S A*, **102**(28):9830–5.

Johnson, D., Davidson, B., Brown, C., Smith, W., and Sidow, A., 2004. Noncoding regulatory sequences of ciona exhibit strong correspondence between evolutionary constraint and functional importance. *Genome Res*, **14**(12):2448–56.

Kent, W., 2002. Blat–the blast-like alignment tool. *Genome Res*, **12**(4):656–64.

Kirouac, M. and Sternberg, P., 2003. cis-regulatory control of three cell fate-specific genes in vulval organogenesis of caenorhabditis elegans and c. briggsae. *Dev Biol*, **257**(1):85–103.

Kusakabe, T., 2005. Regulation and evolution of genes in ascidians. *Zoolog Sci*, **22**(12):1372.

Lee, A., Koh, E., Tay, A., Brenner, S., and Venkatesh, B., 2006. Highly conserved syntenic blocks at the vertebrate hox loci and conserved regulatory elements within and outside hox gene clusters. *Proc Natl Acad Sci U S A*, **103**(18):6994–9.

Lemons, D. and McGinnis, W., 2006. Genomic evolution of hox gene clusters. *Science*, **313**(5795):1918–22.

Livi, C. and Davidson, E., 2007. Regulation of spblimp1/krox1a, an alternatively transcribed isoform expressed in midgut and hindgut of the sea urchin gastrula. *Gene Expr Patterns*, **7**(1-2):1–7.

Loots, G., Locksley, R., Blankespoor, C., Wang, Z., Miller, W., Rubin, E., and Frazer, K., 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, **288**(5463):136–40.

Ludwig, M., Patel, N., and Kreitman, M., 1998. Functional analysis of eve stripe 2 enhancer evolution in drosophila: rules governing conservation and change. *Development*, **125**(5):949–58.

Macdonald, S. and Long, A., 2005. Identifying signatures of selection at the enhancer of split neurogenic gene complex in drosophila. *Mol Biol Evol*, **22**(3):607–19.

Mackenzie, A., Miller, K., and Collinson, J., 2004. Is there a functional link between gene interdigitation and multi-species conservation of synteny blocks? *Bioessays*, **26**(11):1217–24.

Margulies, E., Blanchette, M., Haussler, D., and Green, E., 2003. Identification and characterization of multi-species conserved sequences. *Genome Res*, **13**(12):2507–18.

Margulies, E., Chen, C., and Green, E., 2006. Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. *Trends Genet*, **22**(4):187–93.

Matsumata, M., Uchikawa, M., Kamachi, Y., and Kondoh, H., 2005. Multiple n-cadherin enhancers identified by systematic functional screening indicate its group b1 sox-dependent regulation in neural and placodal development. *Dev Biol*, **286**(2):601–17.

Mayor, C., Brudno, M., Schwartz, J., Poliakov, A., Rubin, E., Frazer, K., Pachter, L., and Dubchak, I., 2000. Vista : visualizing global dna sequence alignments of arbitrary length. *Bioinformatics*, **16**(11):1046–7.

Minokawa, T., Wikramanayake, A., and Davidson, E., 2005. cis-regulatory inputs of the wnt8 gene in the sea urchin endomesoderm network. *Dev Biol*, **288**(2):545–58.

Mortlock, D., Guenther, C., and Kingsley, D., 2003. A general approach for identifying distant regulatory elements applied to the gdf6 gene. *Genome Res*, **13**(9):2069–81.

Mortlock, D., Portnoy, M., Chandler, R., and Green, E., 2004. Comparative sequence analysis of the gdf6 locus reveals a duplicon-mediated chromosomal rearrangement in rodents and rapidly diverging coding and regulatory sequences. *Genomics*, **84**(5):814–23.

Moses, A., Pollard, D., Nix, D., Iyer, V., Li, X., Biggin, M., and Eisen, M., 2006. Large-scale turnover of functional transcription factor binding sites in drosophila. *PLoS Comput Biol*, **2**(10):e130.

Ning, Z., Cox, A., and Mullikin, J., 2001. Ssaha: a fast search method for large dna databases. *Genome Res*, **11**(10):1725–9.

Oda-Ishii, I., Bertrand, V., Matsuo, I., Lemaire, P., and Saiga, H., 2005. Making very similar embryos with divergent genomes: conservation of regulatory mechanisms of otx between the ascidians halocynthia roretzi and ciona intestinalis. *Development*, **132**(7):1663–74.

Pennacchio, L., Ahituv, N., Moses, A., Prabhakar, S., Nobrega, M., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K., *et al.*, 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, **444**(7118):499–502.

Pollard, D., Bergman, C., Stoye, J., Celniker, S., and Eisen, M., 2004. Benchmarking tools for the alignment of functional noncoding dna. *BMC Bioinformatics*, **5**:6.

Pollard, D., Moses, A., Iyer, V., and Eisen, M., 2006. Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. *BMC Bioinformatics*, **7**:376.

Portnoy, M., McDermott, K., Antonellis, A., Margulies, E., Prasad, A., Kingsley, D., Green, E., and Mortlock, D., 2005. Detection of potential gdf6 regulatory elements by multispecies sequence comparisons and identification of a skeletal joint enhancer. *Genomics*, **86**(3):295–305.

Poulin, F., Nobrega, M., Plajzer-Frick, I., Holt, A., Afzal, V., Rubin, E., and Pennacchio, L., 2005. In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics*, **85**(6):774–81.

Prabhakar, S., Poulin, F., Shoukry, M., Afzal, V., Rubin, E., Couronne, O., and Pennacchio, L., 2006. Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res*, **16**(7):855–63.

Ransick, A. and Davidson, E., 2006. cis-regulatory processing of notch signaling input to the sea urchin glial cells missing gene during mesoderm specification. *Dev Biol*, **297**(2):587–602.

Revilla i Domingo, R., Minokawa, T., and Davidson, E., 2004. R11: a cis-regulatory node of the sea urchin embryo gene network that controls early expression of spdelta in micromeres. *Dev Biol*, **274**(2):438–51.

Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., Green, E., Hardison, R., and Miller, W., 2003a. Multipipmaker and supporting tools: Alignments and analysis of multiple genomic dna sequences. *Nucleic Acids Res*, **31**(13):3518–24.

Schwartz, S., Kent, W., Smit, A., Zhang, Z., Baertsch, R., Hardison, R., Haussler, D., and Miller, W., 2003b. Human-mouse alignments with blastz. *Genome Res*, **13**(1):103–7.

Schwartz, S., Zhang, Z., Frazer, K., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W., 2000. Pipmaker–a web server for aligning two genomic dna sequences. *Genome Res*, **10**(4):577–86.

Settle, S., Rountree, R., Sinha, A., Thacker, A., Higgins, K., and Kingsley, D., 2003. Multiple joint and skeletal patterning defects caused by single and double mutations in the mouse gdf6 and gdf5 genes. *Dev Biol*, **254**(1):116–30.

Sinha, S., Schroeder, M., Unnerstall, U., Gaul, U., and Siggia, E., 2004. Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in drosophila. *BMC Bioinformatics*, **5**:129.

Sonnhammer, E. and Durbin, R., 1995. A dot-matrix program with dynamic threshold control suited for genomic dna and protein sequence analysis. *Gene*, **167**(1-2):GC1–10.

Symula, D., Frazer, K., Ueda, Y., Denefle, P., Stevens, M., Wang, Z., Locksley, R., and Rubin, E., 1999. Functional screening of an asthma qtl in yac transgenic mice. *Nat Genet*, **23**(2):241–4.

Takahashi, H., Mitani, Y., Satoh, G., and Satoh, N., 1999. Evolutionary alterations of the minimal promoter for notochord-specific brachyury expression in ascidian embryos. *Development*, **126**(17):3725–34.

Teng, Y., Girard, L., Ferreira, H., Sternberg, P., and Emmons, S., 2004. Dissection of cis-regulatory elements in the c. elegans hox gene egl-5 promoter. *Dev Biol*, **276**(2):476–92.

Thacker, C., Marra, M., Jones, A., Baillie, D., and Rose, A., 1999. Functional genomics in caenorhabditis elegans: An approach involving comparisons of sequences from related nematodes. *Genome Res*, **9**(4):348–59.

Thompson, J., Higgins, D., and Gibson, T., 1994. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**(22):4673–80.

Uchikawa, M., Ishida, Y., Takemoto, T., Kamachi, Y., and Kondoh, H., 2003. Functional analysis of chicken sox2 enhancers highlights an array of diverse regulatory elements that are conserved in mammals. *Dev Cell*, **4**(4):509–19.

Venkatesh, B. and Yap, W., 2005. Comparative genomics using fugu: a tool for the identification of conserved vertebrate cis-regulatory elements. *Bioessays*, **27**(1):100–7.

Waterston, R., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.*, 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**(6915):520–62.

Wittkopp, P., 2006. Evolution of cis-regulatory sequence and function in diptera. *Heredity*, **97**(3):139–47.

Yagi, K., Satou, Y., and Satoh, N., 2004. A zinc finger transcription factor, zicl, is a direct activator of brachyury in the notochord specification of ciona intestinalis. *Development*, **131**(6):1279–88.

Yuh, C., Brown, C., Livi, C., Rowen, L., Clarke, P., and Davidson, E., 2002. Patchy interspecific sequence similarities efficiently identify positive cis-regulatory elements in the sea urchin. *Dev Biol*, **246**(1):148–61.

Chapter 3

Paircomp, FamilyRelationsII and Cartwheel: Tools for Interspecific Sequence Comparison

This work is already published.[1]

3.1 Background

Comparative sequence analysis is fast becoming a standard method for discovering cis- regulatory modules (Cooper and Sidow, 2003). The technique relies on the signatures of conservation left by functional genomic regions as the background sequence evolves. It is often the only way to computationally discover cis-regulatory modules in animal genomes when definite knowledge of upstream regulators is lacking, and it can serve as an excellent complement to experimental techniques.

Paircomp, FamilyRelationsII (FRII), and Cartwheel are an integrated system for comparing two BAC-sized (100kb) genomic sequences, viewing the comparison, manipulating thresholds and views, and extracting the results. These tools and their predecessors, seqcomp and FamilyRelations, have been used extensively in the years since we first made them available (Brown et al., 2002). However, the addition of Cartwheel, a Web server system for performing, storing, and revisiting analyses, makes this combined toolkit considerably more useful to the experimental biologist.

The first analysis done with FamilyRelations was a comparison of the otx region between two sea urchins; 11 of the 17 conserved blocks were shown to drive expression of a reporter (Yuh et al., 2002). Kirouac and Sternberg (Kirouac and Sternberg, 2003) showed that features conserved between C. elegans and C. briggsae encoded functional regulatory regions. Romano and Wray (Romano and Wray, 2003) used FamilyRelations to show that primary sequence identity was conserved in only part of the previously identified endo16 cis-regulatory region, when the *L. variegatus* sequence was used as a partner to the S. purpuratus sequence. Leung et al. (Leung et al., 2004) used FRII to analyze regions in which NFKB bound to verify that the regions were conserved between mouse and human. And, most recently, Revilla-i-Domingo et al. (Revilla i Domingo et al., 2004) identified a small conserved region in the delta genomic locus as a cis-regulatory element responsible for localized expression of delta in S. purpuratus. Similar analyses of the regulation of gatae, krox, wnt8, brachyury, tbrain, foxa and deadringer in S. purpuratus are forthcoming from this lab. While most published use of FRII and Cartwheel has been in sea urchins and nematodes, users have reported that the tools accurately identify regulatory regions in vertebrates and plants.

FRII and Cartwheel are specialized for identifying conservation within relatively small genomic regions, and can be used for comparing BAC sequences between organisms for which no whole genome assembly exists (e.g. S. purpuratus/*L. variegatus*). The exhaustive "dot-plot"-style search algorithm used (described below) assumes nothing about the relative positioning or orientation of regulatory regions and can be used to detect rearrangements that might be missed by a global alignment algorithm (see e.g. (Kirouac and Sternberg, 2003)). Because of these features, FRII and Cartwheel are particularly useful in

[1] Brown CT, Xie Y, Davidson EH, Cameron RA. *Paircomp, FamilyRelationsII and Cartwheel: tools for interspecific sequence comparison.* **BMC Bioinformatics**. 2005 Mar 24;6:70.

targeted searches for regulatory regions.

In this paper, we present these effective tools for comparative sequence analysis to the wider biological community.

3.2 Implementation

Paircomp is a program for doing windowed comparisons of two sequences. It is an expanded reimplementation of the seqcomp program (Brown et al., 2002). Paircomp contains several algorithms for doing exhaustive fixed-width-window sequence comparisons, optimized for different parameters. The default algorithm uses a sliding window to do a "rolling comparison" and runs in time O(NxM) for two sequences of lengths N and M. Paircomp is written in C++ and has a Python interface.

FamilyRelationsII (FRII) is a graphical viewer for sequence analyses. It is a C++ reimplementation of the original Java/Jython FamilyRelations (Brown et al., 2002). FRII uses the cross-platform FLTK windowing toolkit to present a common interface on Windows, Mac OS X, and Linux/X11.

Cartwheel is a server-side system that presents a uniform interface for job coordination and execution. It has several components, including a Web interface through which users can establish analyses; a remote interface for programs to retrieve analysis data; and a batch job queueing system based on a method of parallel processing known as a Linda tuple space. All of the components are built on top of a PostgreSQL database. Cartwheel is written in Python and provides libraries in Python, Java, and C++ for remote access.

A technical history of the design decisions made in the implementation of these tools has been published online (http://www.pyzine.com/Issue006/, article "Python in Bioinformatics").

Availability FRII is freely available for download in a binary distribution for Mac OS X and Windows at http://family.caltech.edu/; FRII will also run under most UNIX distributions but must be compiled individually. The Center for Computational Regulatory Genomics at Caltech maintains a public Cartwheel server at http://woodward.caltech.edu/. A tutorial for FRII is available at http://family.caltech.edu/tutorial/, and an example homework assignment for an undergraduate class is also at that Web site. The source code for paircomp, FRII and Cartwheel and all their components is freely available under the L/GPL through the above Web sites. Paircomp, FamilyRelationsII and Cartwheel are Copyright 2001- 2004 the California Institute of Technology.

3.3 Results and discussion

3.3.1 Paircomp

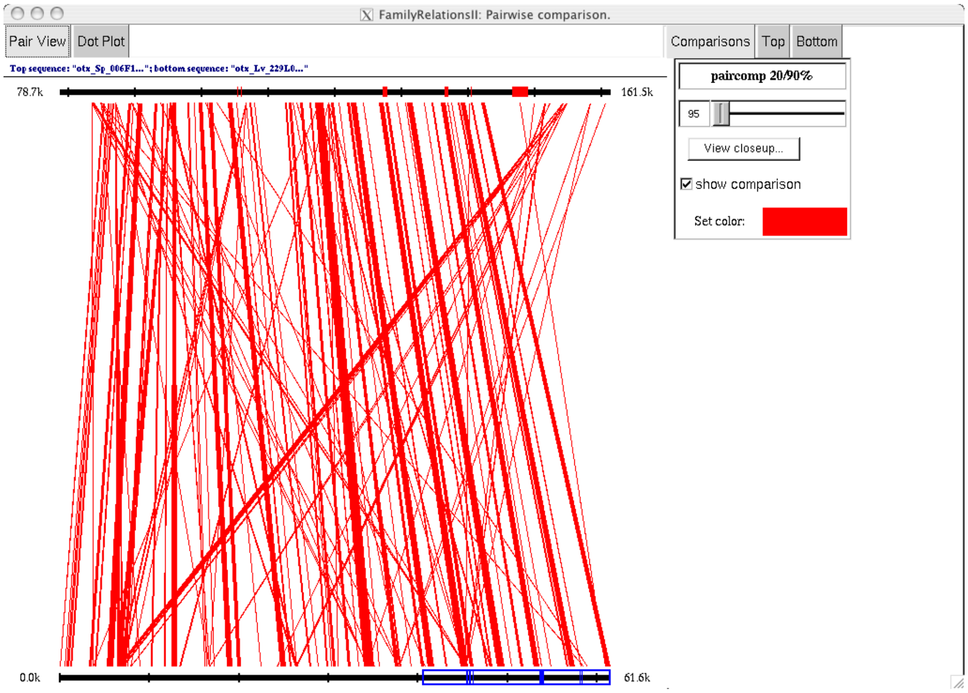
Several different classes of algorithms are available for comparing two genomic sequences. Windowed comparisons do an exhaustive comparison of two sequences with a fixed-width window, and record strict (ungapped) sequence identity within that window (Sonnhammer and Durbin, 1995, Brown et al., 2002). Local alignment algorithms such as BLAST search for common "words" of DNA in a pair of sequences and build a gapped alignment around these words (Altschul et al., 1990). These gapped alignments are often scored by overall length, so that e.g. a 500bp match at 90% is ranked higher than a 200bp match at 90%. Global alignment algorithms such as AVID (Bray et al., 2003) and LAGAN (Brudno et al., 2003) seek to build a start-to-end gapped alignment of syntenic genomic regions. Windowed comparisons and local alignment algorithms usually search for matches in both forward and reverse complement directions, while global alignment algorithms typically try to build an alignment without inversions. Implementations of all three strategies for genomic comparisons have been publicly available for some time: Dotter and seqcomp implement windowed comparisons (Sonnhammer and Durbin, 1995, Brown et al., 2002); PipMaker uses a local alignment algorithm, blastz (Elnitski et al., 2002, Schwartz et al., 2000); and Vista relies on a global alignment generated by AVID (Frazer et al., 2004). All three comparison strategies have been successful at finding regulatory regions (Yi et al., 1991, Cooper and Sidow, 2003).

Of the three general classes of algorithms, we chose to use windowed comparisons in our search for cis-regulatory modules. Our decision was based on several criteria. First, these comparisons report matches based solely on strict sequence identity with no gapping, unlike alignment algorithms. This is a good ab initio requirement when comparing sequences in search of cis-regulatory modules, whose evolution is still poorly understood; in particular, binding sites could be sensitive to indels, which are somewhat elided in gapped alignments. Moreover, we had no a priori expectation for the locations, sizes, or degrees of similarity of conserved regions, necessitating an exhaustive search strategy that did not bias scores based on the length or position of matches. And, finally, from a user-interface perspective the parameters for paircomp windowsize and threshold are simple and intuitively linked to the results. Our success with this basic approach means that we have not needed to move to alternative algorithms.

Paircomp is a standalone program that executes windowed comparisons (see Methods). It searches for matches in both the forward and reverse complement directions. Paircomp runs within Cartwheel; the results are stored in a database and communicated to FRII.

3.3.2 Cartwheel

Cartwheel is a Web site through which analyses are executed and from which analyses are loaded into FamilyRelationsII. It provides an easy-to-use interface through which to establish a set of analyses on a pair of sequences. Cartwheel also allows the annotation of sequences with a variety of features; features can be uploaded to Cartwheel in the standard GFF format. A tutorial for setting up pairwise comparisons is online at http://family.caltech.edu/tutorial/.



FamilyRelationsII: Pairwise comparison.

Top sequence: "otx_Sp_006F1..."; bottom sequence: "otx_Lv_229L0..."

paircomp 20/90%

95

☑ show comparison

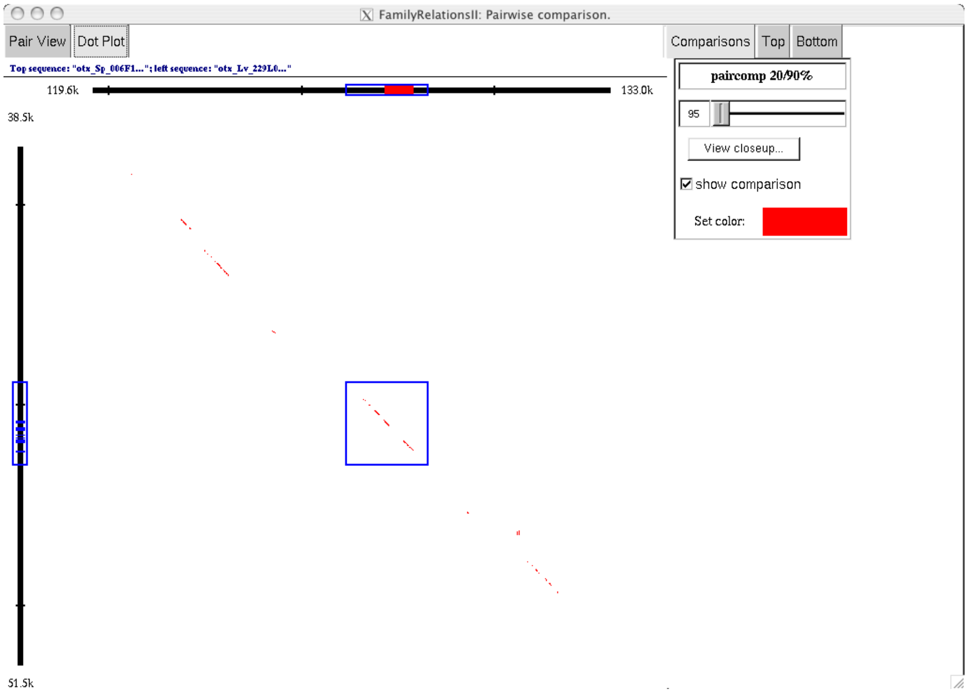
Figure 3.1: A paircomp comparison of the otx gene locus from S. purpuratus (top) with L. variegatus (bottom). We used paircomp to compare all 20bp subsequences from a 160kb S. purpuratus BAC with a 62kb *L. variegatus* BAC; those 20bp subsequences with a exact match of 19/20 or 20/20 bases are connected with a red line. Only the 80kb surrounding the otx gene is shown on the top. Matches to the known S. purpuratus cDNA sequence are shown in red on the top sequence, and TBLASTX matches in L. variegatus to the same cDNA sequence are shown in blue on the bottom sequence. The *L. variegatus* genomic sequence does not extend to cover the 3' region of the coding sequence. On the top of the view are tabs to switch between the "pair view" (shown) and the "dot plot" view (see Figure 3.2). On the right side of the view are control buttons that allow the user to change both the color and the threshold at which matches are displayed. The user can also view a closeup of a region by selecting the region on the sequence (e.g. as on the bottom sequence, where a region from 40kb to 61.6kb is selected) and then pressing the "View closeup" button. An example closeup view is shown in Figure 3.3.

3.3.3 FamilyRelationsII

FamilyRelationsII, or FRII, displays comparisons of BAC-sized genomic sequences of lengths 100kb. It is a graphical program that runs directly from a desktop and loads data from the Cartwheel server. From within FRII, users can zoom in to look more closely at features, alter scoring thresholds for comparisons, change the color of features, and turn on or off the display of specific analyses. FRII can also display closeup views of comparisons and alignments against DNA and protein sequence.

Figure 3.1 shows the main FRII view of a comparison between the otx locus in S. purpuratus and *L. variegatus*, two sea urchins that diverged approx. 50 mya. The genomic sequences were obtained from BAC libraries as described in (Yuh et al., 2002). In the case of S. purpuratus, the BAC contains the entire otx coding region; the L. variegatus sequence contains only the 5' region of the gene, and not the final exon.

The comparison shown is a paircomp comparison performed with a 20bp window at 90% and then



FamilyRelationsII: Pairwise comparison.

Top sequence: "otx_Sp_006F1..."; left sequence: "otx_Lv_229L0..."

paircomp 20/90%

95

☑ show comparison

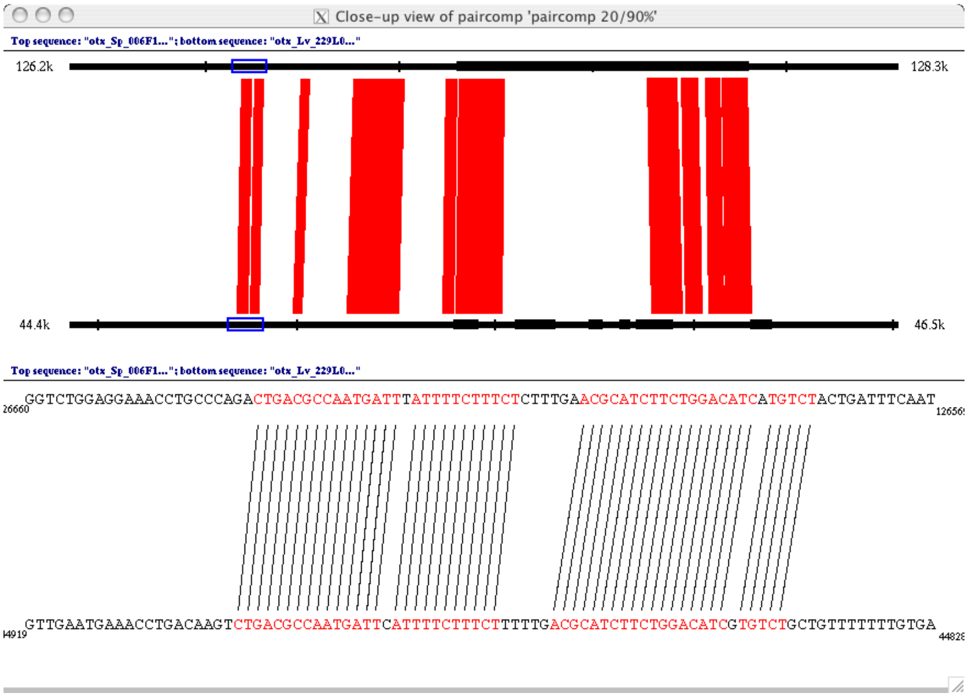
Figure 3.2: A "dot-plot" style view of a subregion of the otx comparison (see Figure 3.1). The top sequence is a zoomed-in view of the otx genomic region from *S. purpuratus*, as in Figure 3.1; the region runs from 119.6kb to 133.0kb. The side sequence is a zoomed-in view of the orthologous region from *L. variegatus*, running from 38.5kb to 51.5kb. The region surrounding the first exon (in red) of the sp-otx transcript is selected on the top (*S. purpuratus*) sequence, and the corresponding TBLASTX matches are highlighted on the left (*L. variegatus*) sequence in blue. The selection box in the center of the view contains the paircomp matches in this region, showing only 20bp matches that match at 19/20 or 20/20 (corresponding to a 95% threshold). A closeup view of this region, showing the DNA sequence of the two regions with the corresponding matches, is shown in Figure 3.

displayed at a 95% threshold. The general colinearity of the matches suggests that the majority of the similar regions are conserved with respect to size, orientation, and relative distance from the exons. This colinearity is typical of conserved features in our comparisons. The diagonal lines crossing the comparison often identify low complexity regions such as simple sequence repeats present throughout both genomic regions. This pairwise mapping view is one of the two large-scale views in FRII; the other large-scale view is a dot-plot view, shown in Figure 3.2.

Figure 3.2 shows a dot-plot view of an expanded region of the comparison, centered on the first exon of the otx transcript. In addition to the exon itself, there is patchy conservation throughout the region; again, this is typical of many comparisons. This view also shows that all of the elements are collinear on scales of 10kb.

In both the dot-plot and pairwise mapping view, multiple comparisons done with different parameters can be displayed in different colors. The threshold for the matches shown can be adjusted until the desired view is obtained, and sequence can be exported from any of the views via a pop-up menu.

Once a threshold is chosen, the user can expand the view of a particular region. Figure 3.3 shows a



Top sequence: "otx_Sp_006F1..."; bottom sequence: "otx_Lv_229L0..."

Top sequence: "otx_Sp_006F1..."; bottom sequence: "otx_Lv_229L0..."

```
GGTCTGGAGGAAACCTGCCCAGACTGACGCCAATGATTTATTTTCTTTCTCTTTGAACGCATCTTCTGGACATCATGTCTACTGATTTCAAT
GTTGAATGAAACCTGACAAGTCTGACGCCAATGATTCATTTTCTTTCTTTTTGACGCATCTTCTGGACATCGTGTCTGCTGTTTTTTTGTGA
```

Figure 3.3: A closeup view of the paircomp comparison of the genomic sequence surrounding the first exon of otx in *S. purpuratus* (top sequence) and *L. variegatus* (bottom sequence). The top half of the closeup view shows orthologous 2kb genomic regions (126.2kb - 128.3kb in the *S. purpuratus* BAC, 44.4kb - 46.5kb in the L. variegatus BAC). Matches of 19/20 or 20/20 bases are drawn in red between the sequences, and the exon matches from Figure 3.2 are shown in black on the sequence lines. The bottom half of the closeup view shows the part of the sequence selected in blue on the top half of the view. Lines are drawn in black between individual matching bases, and the matching bases are colored in red. Note that both blocks shown match at 19/20 because of the single mismatch in the middle of the blocks.

closeup view of the region outlined in blue in Figure 3.2. The sequence shown in Figure 3.3 is a small patch of conservation upstream of the first exon, displayed at a 19/20 threshold. Here the user scans along the sequence and visually compares both the boundaries of the matches and the complexity of the sequence. Sequences are directly exported to other applications via the "paste" buffer.

FRII also performs searches for motifs using the IUPAC notation in which e.g. W represents A or T. This feature allows users to search for matches to known "consensus" binding sites for transcription factors. Searches are either stored on the Cartwheel server and displayed as individual features on FRII views, or executed directly in FRII. One particularly convenient feature is the ability to ask for motifs that have mismatches in up to 5 positions; this lets users search for weaker matches to known consensi.

3.3.4 Other analyses

FRII displays a variety of analyses. In addition to paircomp windowed comparisons, FRII displays and manipulates Vista-style comparisons, BLAST and blastz comparisons, BLAST database searches, cDNA and protein comparisons, and the results of several different gene finders (genscan, geneid, and hmmgene

(Burge and Karlin, 1997, Parra et al., 2000, Krogh, 2000)). All of these analyses are executed directly on the Cartwheel server, excepting only Vista comparisons using the (default) AVID alignment program. The data for Vista comparisons must be uploaded from the results returned by the Vista Web site; however, Vista-style comparisons with the LAGAN global alignment tool are executed directly on Cartwheel.

3.3.5 Discovering and analyzing regulatory regions

We and others have successfully used paircomp, FRII, and Cartwheel to discover a number of regulatory regions (see Introduction). Once we have a pair of genomic regions to compare, the steps we follow are essentially invariant from region to region:

1. We set up two to three paircomp analyses at the following windowsizes and thresholds: 10bp/90%; 20bp/80%; 50bp/60%.

2. We match the cDNA or protein of interest against both regions, to determine where the coding regions lie.

3. We also compare the RefSeq database from NCBI against both regions, to find other genes in the region.

4. We load these analyses into FRII and zoom in to a view that includes as much intergenic sequence around the gene as is possible without also including other genes. We then adjust the thresholds on the 20bp and 50bp analyses until we obtain a roughly collinear pattern of conserved blocks. Typical values for these thresholds are 80-100% for a 20bp windowed comparison, and 60-80% for a 50bp windowed comparison.

5. We use the closeup view to extract the conserved blocks, and design PCR primers to isolate all of the contiguous blocks of conserved sequence. We then individually subclone or fuse them into a GFP reporter construct together with a basal promoter. These constructs are then introduced into the sea urchin by microinjection and analyzed for appropriate spatiotemporal expression.

In our experience, we have always been able to identify the relevant enhancer elements using this procedure. A similar procedure in which putatively negative elements are fused with a ubiquitous driver of expression often identifies necessary repressive elements. Also note that one caveat of these procedures is that for some genes, e.g. transcription factors, there are often many regions that appear to do nothing. These may be regulatory regions that affect expression at times or in places that are not under consideration, or could be other genomic features not relevant to gene regulation.

3.4 Conclusions

Paircomp, FamilyRelationsII, and Cartwheel are an effective, easy-to-use set of tools for analyzing conservation in BAC-sized genomic regions. Over 100 people are currently using them, and they have been effective in finding regulatory regions in a variety of organisms. In this paper we have described the tools and provided an introduction for biologists who wish to use them.

3.5 Appendix

3.5.1 Availability and requirements

See Implementation, above, for information on server-side software.

Project name: FamilyRelationsII

Project home page: http://family.caltech.edu/

Operating systems: Mac OS X, Windows NT/XP, UNIX/Linux (X Windows)

Programming language: C++

License: GPL/LGPL

No restrictions placed on use.

3.5.2 Author contributions

CTB designed and implemented the majority of the functionality described. YX implemented a significant portion of the XML-RPC functionality used for client-server interaction. EHD laid out the design requirements, aided in writing the paper, and supervised the development of FRII. RAC is responsible for running the servers and did the majority of bug testing, and also contributed to the paper.

3.5.3 Acknowledgements

Tristan De Buysscher and Madeleine Price, under the supervision of Dr. Barbara Wold, developed the original seqcomp and contributed to FamilyRelations. Ramon Cendejas and Kevin Berney aided in the development of features and helped exercise the Cartwheel server; a complete list of contributors to FamilyRelationsII and Cartwheel can be found on the Cartwheel Web site, under Developers. We especially thank Carolina Livi, Pei-Yun Lee, Dr. Ellen Rothenberg and Dr. Erich Schwarz for extensive user-interface testing over the years. In addition, Dr. Ellen Rothenberg and Dr. Erich Schwarz both contributed significantly to discussions of new features; in addition, Sagar Damle, Tracy Teal and Dr. Erich Schwarz gave many helpful comments on this paper. We also thank two anonymous reviewers for their comments. CTB is supported by National Institutes of Health Grant GM61005, and the Beckman Institute Center for Computational Regulatory Genomics is supported by National Institutes of Health Grant RR15044.

Bibliography

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D., 1990. Basic local alignment search tool. *J Mol Biol*, **215**(3):403–10.

Bray, N., Dubchak, I., and Pachter, L., 2003. Avid: A global alignment program. *Genome Res*, **13**(1):97–102.

Brown, C., Rust, A., Clarke, P., Pan, Z., Schilstra, M., Buysscher, T. D., Griffin, G., Wold, B., Cameron, R., Davidson, E., *et al.*, 2002. New computational approaches for analysis of cis-regulatory networks. *Dev Biol*, **246**(1):86–102.

Brudno, M., Do, C., Cooper, G., Kim, M., Davydov, E., Green, E., Sidow, A., and Batzoglou, S., 2003. Lagan and multi-lagan: efficient tools for large-scale multiple alignment of genomic dna. *Genome Res*, **13**(4):721–31.

Burge, C. and Karlin, S., 1997. Prediction of complete gene structures in human genomic dna. *J Mol Biol*, **268**(1):78–94.

Cooper, G. and Sidow, A., 2003. Genomic regulatory regions: insights from comparative sequence analysis. *Curr Opin Genet Dev*, **13**(6):604–10.

Elnitski, L., Riemer, C., Petrykowska, H., Florea, L., Schwartz, S., Miller, W., and Hardison, R., 2002. Piptools: a computational toolkit to annotate and analyze pairwise comparisons of genomic sequences. *Genomics*, **80**(6):681–90.

Frazer, K., Pachter, L., Poliakov, A., Rubin, E., and Dubchak, I., 2004. Vista: computational tools for comparative genomics. *Nucleic Acids Res*, **32**(Web Server issue):W273–9.

Kirouac, M. and Sternberg, P., 2003. cis-regulatory control of three cell fate-specific genes in vulval organogenesis of caenorhabditis elegans and c. briggsae. *Dev Biol*, **257**(1):85–103.

Krogh, A., 2000. Using database matches with for hmmgene for automated gene detection in drosophila. *Genome Res*, **10**(4):523–8.

Leung, T., Hoffmann, A., and Baltimore, D., 2004. One nucleotide in a kappab site can determine cofactor specificity for nf-kappab dimers. *Cell*, **118**(4):453–64.

Parra, G., Blanco, E., and Guigo, R., 2000. Geneid in drosophila. *Genome Res*, **10**(4):511–5.

Revilla i Domingo, R., Minokawa, T., and Davidson, E., 2004. R11: a cis-regulatory node of the sea urchin embryo gene network that controls early expression of spdelta in micromeres. *Dev Biol*, **274**(2):438–51.

Romano, L. and Wray, G., 2003. Conservation of endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-acting components of transcriptional regulation. *Development*, **130**(17):4187–99.

Schwartz, S., Zhang, Z., Frazer, K., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W., 2000. Pipmaker–a web server for aligning two genomic dna sequences. *Genome Res*, **10**(4):577–86.

Sonnhammer, E. and Durbin, R., 1995. A dot-matrix program with dynamic threshold control suited for genomic dna and protein sequence analysis. *Gene*, **167**(1-2):GC1–10.

Yi, T., Walsh, K., and Schimmel, P., 1991. Rabbit muscle creatine kinase: genomic cloning, sequencing, and analysis of upstream sequences important for expression in myocytes. *Nucleic Acids Res*, **19**(11):3027–33.

Yuh, C., Brown, C., Livi, C., Rowen, L., Clarke, P., and Davidson, E., 2002. Patchy interspecific sequence similarities efficiently identify positive cis-regulatory elements in the sea urchin. *Dev Biol*, **246**(1):148–61.

Chapter 4

A Logic Diagram for *cyIIIa*

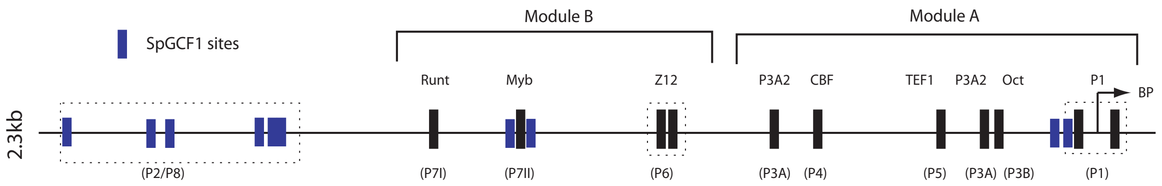


Figure 4.1: The *cyIIIa cis*-regulatory region. The horizontal line represents the approximately 2300 bases of DNA immediately 5' of the transcription start site of *cyIIIa*. This region includes the basal promoter (BP) of the *cyIIIa* gene. The black and blue rectangles represent transcription factor binding sites originally identified by (Calzone et al., 1988). The blue rectangles mark binding sites for SpGCF1, a ubiquitous positive-acting DNA looping factor commonly found in sea urchin regulatory regions (Zeller et al., 1995c). The black rectangles are binding sites for 8 different species of transcriptional regulators, whose identities are mostly known; see text for details. The original designations of these sites (P1 through P7) are taken from (Calzone et al., 1988) and are indicated below the rectangles in parentheses. The division of this region into two modules of independent function by (Kirchhamer and Davidson, 1996) is indicated above the sequence line by the brackets.

4.1 Introduction

The *cyIIIa* gene expresses a cytoskeletal actin during larval embryogenesis in the California purple sea urchin, *S. purpuratus*. Transcription of *cyIIIa* begins at 9-10 hpf, during late cleavage, and *cyIIIa* is expressed only in the aboral ectoderm of the developing embryo (Shott et al., 1984, Angerer and Davidson, 1984, Cox et al., 1986). A series of studies has established that both the temporal and spatial prevalence of the *cyIIIa* transcript are predominantly due to regulation of the rate of transcriptional initiation (summarized in (Lee et al., 1992)).

The 2.3kb of genomic DNA immediately adjacent to the 5' end of the transcription start site of *cyIIIa* drives reporter constructs in the same spatiotemporal pattern as the endogenous *cyIIIa* transcript, i.e. in the aboral ectoderm starting at 9 hours (Flytzanis et al., 1987, Hough-Evans et al., 1988). Thus this genomic DNA is a *cis*-regulatory region containing all of information necessary for the correct spatiotemporal regulation of *cyIIIa*.

(Calzone et al., 1988) used 24 hour nuclear extract to identify 15 binding sites in this *cis*-regulatory region. 9 distinct transcription factors bind to these 15 sites. The organization of these sites, as well as results from *in vivo* titration assays using fragments of the regulatory region, suggested that the *cyIIIa cis*-regulatory region possesses a modular substructure with multiple distinct driver/repressor relationships (Franks et al., 1990, Hough-Evans et al., 1990, Kirchhamer and Davidson, 1996). (Kirchhamer and Davidson, 1996) systematically investigated the roles of these binding sites through site deletion and divided the *cyIIIa cis*-regulatory region into three functional modules, modules A, B, and C.

The *cyIIIa cis*-regulatory system contains binding sites for the SpGCF1 transcription factor throughout all modules. SpGCF1 was sequenced by (Zeller et al., 1995b), and found to be a multimerizing transcription factor that loops DNA *in vitro* (Zeller et al., 1995c). *In vivo* titration with SpGCF1 binding sites showed that SpGCF1 does not drive time-varying or spatially localized activity: it merely amplifies ex-

Binding site	**Transcription factor**	**Description**
P1	*unknown*	An unidentified factor that binds to two sites surrounding the basal promoter. Driver of module A expression in the ectoderm.
P3A	SpP3A2	A pan-embryonic factor that is post-translationally regulated in response to a redox gradient. SpP3A2 functions to repress *cyIIIa* expression in the oral ectoderm (Calzone et al., 1991, Hoog et al., 1991, Zeller et al., 1995a).
P3B	SpOct-1	A POU domain factor first sequenced by (Char et al., 1993). A positive regulator of *cyIIIa*.
P4	SpCBF-A/B/C *(aka NFY-B/A/C)*	SpCBF-A/B/C is a heterotrimeric transcription factor; all three subunits are required for DNA binding. SpCBF-B was described by (Li et al., 2002). SpCBF-A was described by (Li et al., 2002). SpCBF-C is discussed in this work. SpCBF-A/B/C is a positive regulator of *cyIIIa*.
P5	SpTEF-1	A homolog of DmScalloped (described in this work). A positive regulator of *cyIIIa*.
P6	SpZ12-1	Described by (Wang et al., 1995b). Responsible for repressing *cyIIIa* expression in the skeletogenic mesenchyme.
P7I	SpRunx	Primary driver of the B module, described by (Coffman et al., 1996). Present throughout the embryo.
P7II	SpMyb	A transcriptional repressor that blocks *cyIIIa* expression in the endoderm and the oral ectoderm. Described by (Coffman et al., 1997).

Figure 4.2: Binding sites present in the *cyIIIa cis*-regulatory region and, where known, the factors which bind to them.

isting activity uniformly (Franks et al., 1990). It is thought to be involved in intermodule communication (Zeller et al., 1995c).

A module The A module (originally known as the proximal module) is approximately 1kb long (Figure 4.1) and contains 11 binding sites: 4 binding sites for SpGCF1, and 7 binding sites for five other transcription factors.

The P1 binding factor has not been identified. (Kirchhamer and Davidson, 1996) showed that the two P1 binding sites mediate a relatively weak positive interaction that drives reporter expression throughout the ectoderm.

Two P3A2 sites mediate tissue-specific repression in the A module. Both binding sites are necessary for complete oral repression of the P1 factor's activation of *cyIIIa* transcription throughout the ectoderm: When these sites are removed from module A, reporter constructs drive expression in both the oral and aboral ectoderm at 48 hrs (Kirchhamer and Davidson, 1996). This result confirms the results of an *in vivo* titration assay showing that competing P3A2-binding factors away from these sites *in vivo* leads to ectopic expression of *cyIIIa* in the oral ectoderm (Hough-Evans et al., 1990).

The P3A binding sites are bound by two distinct proteins, SpP3A1 and SpP3A2 (Hoog et al., 1991, Calzone et al., 1991, Zeller et al., 1995a). While SpP3A1 and SpP3A2 recognize similar sites, SpP3A1 made *in vitro* binds to a smaller region of the P3A2 binding site than does SpP3A2 purified from nuclear extract, suggesting that either SpP3A2 regulates only a subset of the sites that SpP3A1 regulates, or the *in vitro* protein is incomplete (Hoog et al., 1991). (Zeller et al., 1995a) used antibodies to quantify SpP3A1 and SpP3A2 in the nuclear compartments of embryos and showed that SpP3A2, while ubiquitous, is almost certainly the P3A-binding factor regulating the transcription of *cyIIIa*. This was confirmed by two additional experiments. First, (Bogarad et al., 1998) showed that a single-chain antibody blocking DNA binding by SpP3A2 expanded endogenous *cyIIIa* expression to the entire ectoderm. Second, (Coffman and Davidson, 2001) injected a VP16-SpP3A2 chimeric protein and showed that it was capable of driving expression throughout the ectoderm. This suggests that the oral-specific negative regulation of *cyIIIa* is due to a localized activation of SpP3A2 binding in the oral ectoderm downstream of a redox asymmetry. Thus, while SpP3A2 is ubiquitous, it may bind to DNA only in the oral ectoderm, where it can serve to repress *cyIIIa* transcription.

The P3B and P4 sites were shown by *in vivo* titration to have an effect on the transcription of *cyIIIa*. Both binding factors were tentatively identified by binding site similarity. P3B is probably bound by SpOct-1, a POU-domain transcription factor that is the ortholog of human Oct-1; the binding site is a canonical Oct factor site, and the timing of *cyIIIa* activation is consonant with the timing of other Oct-responsive genes (Char et al., 1993). P4 is probably bound by CCAAT-binding factor (CBF), as the P4 site is identical to a known CBF site at 13/14 bases (Barberis et al., 1987). SpCBF is a heterotrimer which binds to DNA only when all three subunits (CBF-A, CBF-B, and CBF-C) are present (Li et al., 1992, Li et al., 2002). Both SpOct-1 and SpCBF-A/B/C are highly conserved homologs of transcriptional activators in other organisms, and they are known to play a role in positive regulation of other sea urchin genes. This positive action is consonant with the functions determined *in vivo* titration P3B and P4 binding sites.

Both *in vivo* titration and deletion assays suggest that the P5 binding site mediates a powerful activation function (Franks et al., 1990, Kirchhamer and Davidson, 1996). The P5 binding factor has not been confirmed in its identity, but unpublished data based on amino acid sequence recovered from affinity chromatography of nuclear extract suggests that the P5 binding factor is a member of the TEA/ATTS binding domain family (Xian, Coffman and Davidson).

B module The B module (originally known as the middle module) contains six binding sites: two SpGCF1 sites, and four binding sites for three other transcription factors (Calzone et al., 1988).

The P7I binding site is bound by SpRunx (Coffman et al., 1996). SpRunx is the primary driver of the B module: the Runx binding site is required for module B effect (Coffman et al., 1996). *In vivo* titration and deletion assays have shown that Runx can drive *cyIIIa* expression in most tissues of the embryo, suggesting that Runx protein is ubiquitous, although Runx transcript disappears from the aboral ectoderm by the prism stage (Robertson et al., 2002).

The paired P6 binding sites are bound by SpZ12, a zinc-finger transcription factor with twelve zinc fingers (Wang et al., 1995b, Wang et al., 1995a). The P6 binding sites are responsible for repressing Runx-mediated activation in the descendants of the PMCs (Franks et al., 1990, Wang et al., 1995b).

The P7II binding site is bound by SpMyb (Coffman et al., 1997). SpMyb represses Runx mediated activation in the oral ectoderm and skeletogenic mesenchyme (Coffman et al., 1997).

C module The C module consists of a number of SpGCF1 binding sites, and possesses no spatiotemporally specific activity of its own (Kirchhamer and Davidson, 1996). The C module amplifies the output of the rest of the *cis*-regulatory region indiscriminately (Kirchhamer and Davidson, 1996, Coffman et al., 1996).

Module functions The distinct functions of the A, B, and C modules are well understood at this point. The A module functions as an early driver of aboral ectoderm-specific transcription, responding to both a tissue-specific activator and a transcriptional repressor downstream of an early redox asymmetry. Between 18 and 24 hours, the B module becomes active, joining the A module in driving *cyIIIa*'s transcription in the aboral ectoderm. The spatial regulators for the B module are entirely distinct from the A module, setting up two independent drivers of aboral ectoderm expression. The C module has neither positive nor negative spatiotemporal activity on its own, and merely amplifies the output of the A and B modules.

Despite the substantial amount of previous work on *cyIIIa*'s regulation, a number of issues remain unresolved:

- The identity of the factor binding to the P1 binding site is still unknown.

- The identity of the factor binding to the P5 binding site was tentatively identified as a TEF factor (Xian, Coffman, Yuh, and Davidson, unpublished data). This has not been confirmed, however.
- The CCAAT-binding factor that binds to the P4 binding site consists of three subunits (A, B, and C), two of which (A and B) have been identified in the sea urchin (Li et al., 1992, Li et al., 2002). The third subunit, CBF-C, has not yet been found in the sea urchin, although its presence can be inferred from the functionality of the CCAAT-binding factor (Li et al., 2002).
- The P3B, P4, and P5 sites play no spatial role in *cyIIIa*'s regulation, but are required for the full expression level of cyIIIa.CAT (Franks et al., 1990, Kirchhamer and Davidson, 1996). However, their contributions have not been analyzed quantitatively.
- Module B is not capable of driving expression without the A module (Kirchhamer and Davidson, 1996). This suggests that one or more sites in the A module are necessary for B module function, i.e. at least one site in module A acts as a "communicator" site (Istrail and Davidson, 2005). The site or sites necessary for this interaction have not been determined.

In this work we confirm the identity of the P5 binding binding factor as SpTEF-1; verify the presence of SpCBF-C; and analyze the function of the P3B, P4, and P5 sites in the context of both the A and B modules.

4.2 Methods

All constructs were made from the deltaP8 construct (Kirchhamer and Davidson, 1996).

Binding site mutagenesis and construct preparation Binding site mutagenesis was carried out using overlapping internal tailed primers that contained the mutated site(s). Overlapping PCR fragments containing the mutated sites were first generated from the deltaP8 construct and then combined using PCR fusion. Individual mutated regulatory regions (BA(P3Bm), BA(P4m), and BA(P5m)) were then cloned back into the original construct and reporter gene using the internal BglII/EagI sites in deltaP8.

The P3B site was mutated from GCACCGAATCTCATTTGCATATCCTTTT to GCACCGAATCTC-cgggtCATATCCTTTT; the P4 site was mutated from ATAATGGAAACTCTGATTGGACCACGGTGAA to ATAATGGAAACTCTGgcTaaACCACGGTGAA; and the P5 site was mutated from CATTCATTGTCGC-GACATACTTGTAGT to CATTCATgtgCGCGcacTACTTGTAGT.

The BA.GFP construct was made by excising the *cyIIIa cis*-regulatory region with a double digest using EagI and BglII, and then cloning the resulting fragment into the EpGFPII construct.

The BA(shortX2) construct was made by constructing oligos bearing the above P3B, P4, and P5 sites with overlapping ends, and then using fusion PCR to combine the oligos. Each binding site was placed equidistantly in a 240 bp region. This region was then ligated end-to-end via an NheI site and combined via fusion PCR with the B module and a construct containing the basal promoter and mutated P1 sites from the P1m construct (Kirchhamer and Davidson, 1996) to yield BA(shortX2).

The BA(subst) construct was made by constructing oligos bearing the above P3B, P4, and P5 sites with tailed ends matching GFP coding sequence. Multiple rounds of tailed PCR followed by fusion PCR resulted in a construct containing the P3B, P4, and P5 sites in the same spacing as module A, but with completely replaced inter-site sequence.

Oligo and construct sequences are available upon request from the authors. All constructs were confirmed by sequencing.

Preparing constructs for injection GFP and CAT constructs were generated for injection using PCR from plasmid DNA. Injection constructs containing the B module were generated by using the deltaP8 primer (CCCTATACTCGTAATGTAAAAGGGTTTTGCAGCC) at the 5' (distal) end of the B module for PCR, while constructs containing the A module alone were generated using the prox primer (GCGAGAATCATTCAACCATAATGG) at the 5' end of the A module.

Injection solutions were made as described in (Ransick and Davidson, 2006). Microinjections were performed within 10 minutes of fertilization. Embryos were cultured at 15 degrees, and after hatching were transferred to large 6 well plates.

RNA extraction and QPCR At each time point, 100 embryos were collected and stored in lysis buffer and frozen at -80 degrees. The Sigma-Aldrich GenElute Mammalian Total RNA Miniprep Kit (Sigma RTN-70) was used to extract both genomic DNA and RNA. Half of the resulting eluate was subjected to Ambion TURBO DNA-free (Ambion 1907) to remove genomic DNA, and then reverse transcribed into single-strand cDNA with the ABI N808 TaqMan kit. QPCR on both genomic DNA and cDNA was performed on an ABI 7900 using BioRad SybrGreen mix.

cDNA levels were measured using gene-specific primers for cyIIIa, CAT or GFP, and ubiquitin. DNA incorporation levels were assessed by comparing the amount of the endogenous *cyIIIa* gene in the genomic DNA with the amount of CAT or GFP reporter gene in the genomic DNA. All CAT cDNA values reported here are normalized against genomic incorporation levels as in (Revilla i Domingo et al., 2004).

TEF-1 isolation, *in vitro* expression, and gel shifts SpTEF-1 was isolated from a 50-hr RACE library constructed with the Invitrogen GeneRacer kit (Invitrogen L1500-01) using separate (overlapping) forward and reverse primers designed based on earlier TEF-1 sequencing. The resulting PCR fragments were combined with PCR fusion and then subcloned into the Invitrogen pET102/D-TOPO system following the manufacturer's guidelines. Induction and gel shifts were performed as in (Yuh et al., 2004).

4.3 Results

4.3.1 Time course of *cyIIIa* and *cyIIIa.CAT*

RNA titration assays showed that *cyIIIa* is initially expressed at 7-10 hrs, during late cleavage. We used QPCR primers targeting both the 3' UTR of the *cyIIIa* mRNA and the first intron of the nuclear transcript to measure both transcript abundance and transcription rate throughout the first 60 hours of development.

Previous reporter measurements focused on assays of CAT protein abundance, which is two steps removed from the transcriptional initiation rate. CAT protein may also be particularly stable in the squamous epithelium of gastrula-stage embryos (data not shown). Therefore we injected CAT constructs containing both modules A and B as well as a CAT construct containing only module A, and used QPCR to measure relative RNA abundance (Figure 4.3.1). We confirmed that the reporter construct is activated at the same time as the endogenous *cyIIIa* gene, and verified that the mRNA expression of the reporter construct matches the earlier CAT expression profile.

4.3.2 TEF-1 is the factor that binds to the P5 binding site

Amino acid sequence for the regulator of P5 had been obtained using affinity chromatography on nuclear extract with the P5 site, and a cDNA library screen with degenerate oligos recovered clones containing a sea urchin homolog of vertebrate TEF, a TEA/ATTS domain transcription factor (Xian, Coffman, and Davidson, unpublished). Subsequent full-length sequencing of the clones revealed that the clones had recombined with cytochrome C oxidase. Therefore we designed internal primers and used 5' and 3' RACE to recover a full-length TEF clone from a 50 hr cDNA library (see Methods).

The full-length TEF-1 cDNA reconstructed from sequencing of the RACE product contains a 1.3kb protein coding region with a a 300bp 5' UTR and a 150 bp 3' UTR. The predicted coding sequence encodes a 440 amino acid protein that contains a strongly conserved TEA/ATTS DNA binding domain as well as a conserved protein-protein interaction domain for interacting with a coactivator, TONDU (see figure 4.3.2). When transcribed and translated in *E. coli* (see Methods), the resulting protein binds specifically to an oligo containing the P5 binding site and does not bind to a mutated oligo that also does not interact with nuclear extract (see figure 4.3.2; nuclear extract data not shown).

QPCR with *tef-1* specific primers shows that the *tef-1* transcript is present throughout embryogenesis (data not shown). We also found a strongly conserved homolog of the HsTEF-1 coactivating factor, HsTONDU, using a whole-genome BLAST search. The *tondu* transcript is also present throughout embryogenesis (data not shown). Whole mount *in situ* hybridizations of *tef-1* show that *tef-1* is present in all tissues of the embryo (data not shown); the *tondu* transcript is too short to yield an *in situ* signal.

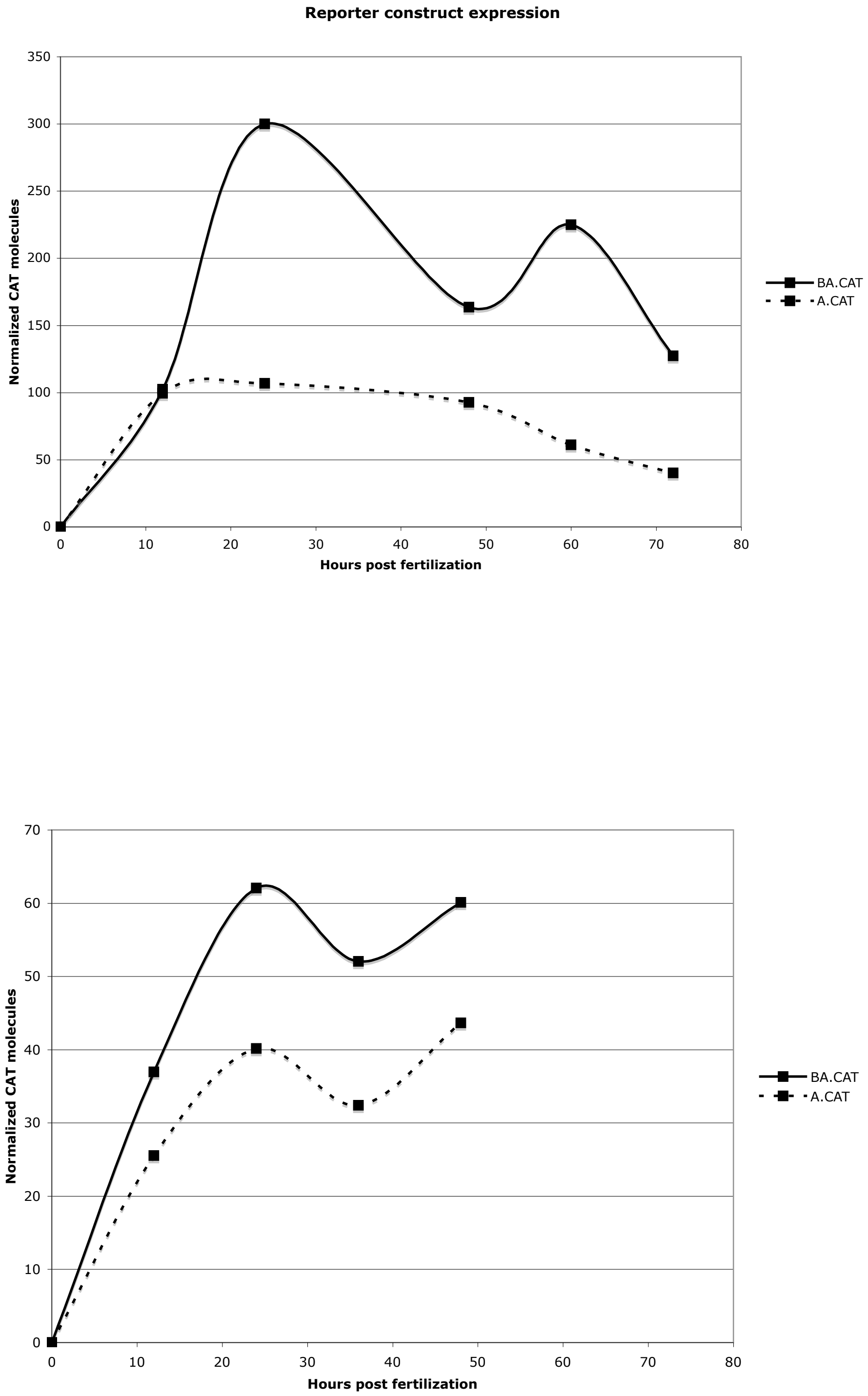


Figure 4.3: Two independent QPCR time courses of CAT reporter construct mRNA prevalence for BA.CAT and A.CAT.

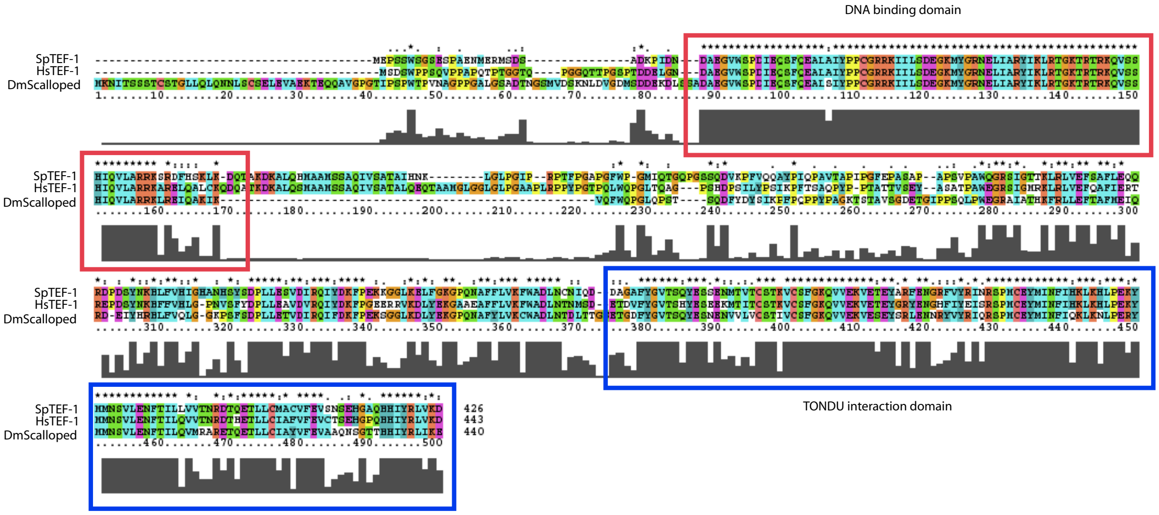


Figure 4.4: *sptef-1* is a TEA/ATTS transcription factor related to *hstef-1* and *dmscalloped*. The alignment between the predicted SpTEF-1, HsTEF-1, and DmScalloped protein sequences shown above demonstrates that SpTEF-1 is a member of the (highly conserved) TEF-1/Scalloped family. The protein sequence contains two strongly conserved domains, a TEA/ATTS DNA binding domain (outlined in red) and a protein-protein interaction domain (outlined in blue) necessary for activation via the TONDU/Vestigial co-activator ((Campbell et al., 1992, Jacquemin et al., 1996, Vaudin et al., 1999)). The domains are nearly identical across all three family members, strongly suggesting that the function of all three proteins is the same.

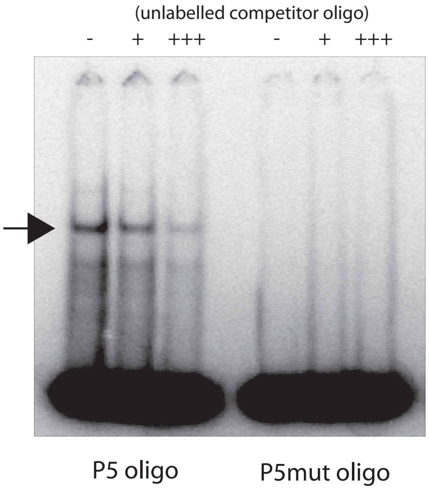


Figure 4.5: *In vitro* transcribed and translated SpTEF-1 binds specifically to the P5 binding site. His-tag-purified SpTEF-1 produced *in vitro* binds to a ^{32}P-labeled oligo containing the P5 binding site from the *cyIIIa cis*-regulatory system (arrow). Binding to the labeled oligo decreases sharply as additional unlabeled oligo is added at either 10-fold (+) or 100-fold (+++) the concentration of the labeled oligo, demonstrating that the interaction is specific. Under the same gel-shift conditions SpTEF-1 does not bind to a mutated sequence in which three of the five core bases in the binding site have been mutated.

4.3.3 *spcbf-c* is transcribed in the embryo

(Li et al., 2002) showed that SpCBF-A and SpCBF-B are present in the sea urchin embryo and function in regulating *spec1*. They could not find SpCBF-C, however. We searched for *sbcbf-c* in the genomic traces with *hscbf-c* using BLAST, and found an excellent single-copy match (data not shown). Using QPCR, we found that a transcript containing this match is initially present maternally and is actively transcribed throughout embryogenesis (data not shown). Thus we know that transcripts encoding all three necessary subunits of SpCBF are present at times during which SpCBF could be regulating *cyIIIa*.

4.3.4 P3B, P4, and P5 binding sites amplify module A output at 30 hours, but not at 50 hours

The early role of the P3B, P4, and P5 binding sites in module A has not been studied quantitatively. To assess the role these sites play in the A module, we mutated each site individually to produce the constructs A(P3B).CAT, A(P4m).CAT, and A(P5m).CAT (shown in figure 4.3.4). We verified via gel shift that the mutated sites no longer bind to any proteins in 24 hour crude nuclear extract (data not shown). We then measured the activity of each construct relative to the original A.CAT construct at both 30 and 50 hours by QPCR (figure 4.3.4).

The results show that at 30 hours the A module depends on all three sites for its normal level of expression, while at 50 hours the sites no longer contribute to the A module output. However, (Franks et al., 1990) showed with *in vivo* titration that all of these sites were important for proper late expression, and (Kirchhamer and Davidson, 1996) showed that a full-length construct missing the P5 site drives reporter expression at a considerably lower level than normal at the gastrula stage (50 hours). This suggests that these sites may also play an important role in the B module.

4.3.5 P3B, P4, and P5 binding sites are required for module B function at 50 hours

Module B function requires one or more of the sites in module A, because module B alone does not drive transcription. The P3B, P4, and P5 sites are likely candidates for this function because the P1 and P3A sites are not necessary for module B function (Kirchhamer and Davidson, 1996). Moreover, *in vivo* titration assays and construct deletion experiments show that the P3B, P4, and P5 sites mediate powerful positive functions in gastrula-stage embryos (Franks et al., 1990, Kirchhamer and Davidson, 1996). This loss of expression is consonant with a loss of module B output resulting from the deletion of a site required for module B function, further suggesting that one or more of these three sites is necessary for module B function.

We mutated each site to abolish the binding interaction in constructs containing both the B and A modules and the A module only (see figure 4.3.4) and verified via gel shift with 24 nuclear extract that

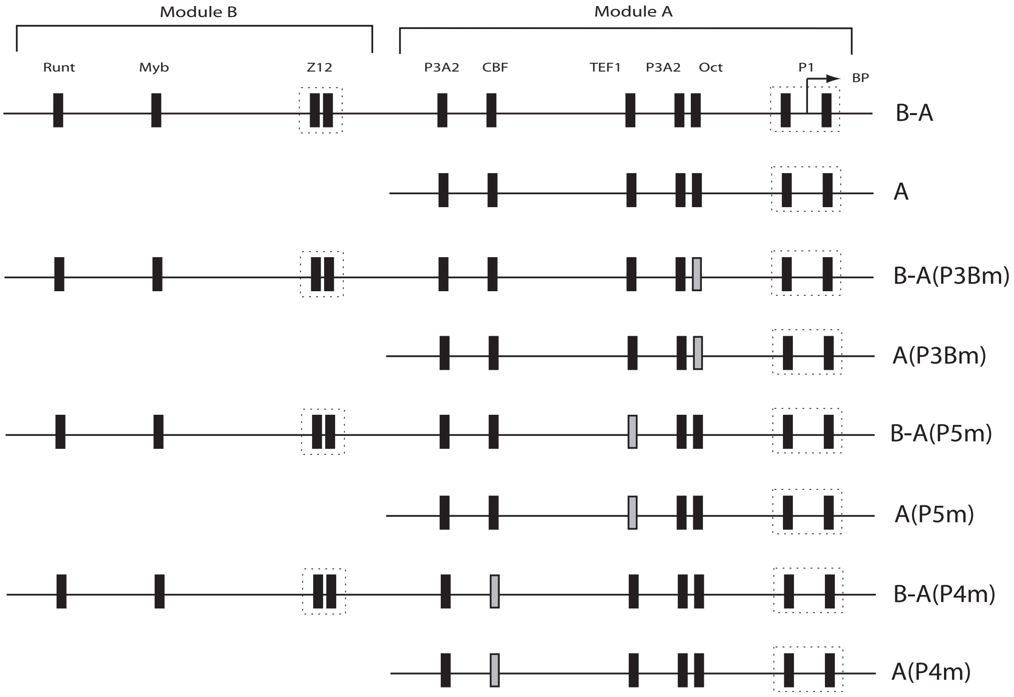


Figure 4.6: A and B module mutant constructs. The binding sites found in (Calzone et al., 1988) are represented in black, labeled with the names of the binding factors. Each pair of BA and A constructs bears a single mutated site (in grey). The order and spacing of sites and the remaining sequence in both modules is otherwise preserved.

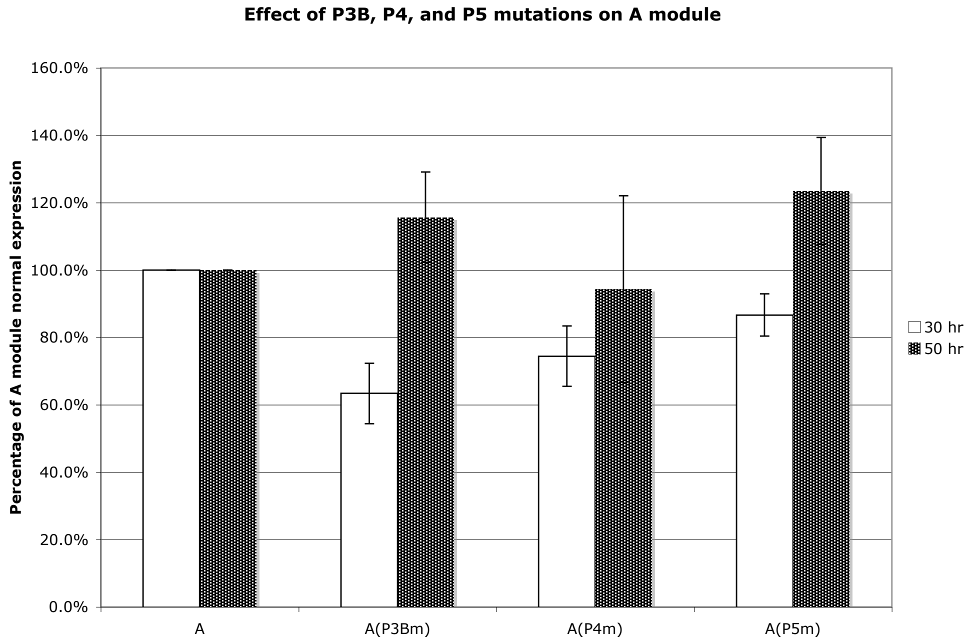


Figure 4.7: QPCR measurements of CAT reporter construct activity for A.CAT, A(P3Bm).CAT, A(P4m).CAT, and A(P5m).CAT at 30 hours (early gastrula) and 50 hours (late gastrula). Expression levels are corrected for transgene incorporation (see Methods) and are shown as percent of normal A.CAT expression measured at 30 or 50 hours in the same experiment. Error bars represent standard deviation across three experiments (30 hour) and two experiments (50 hour).

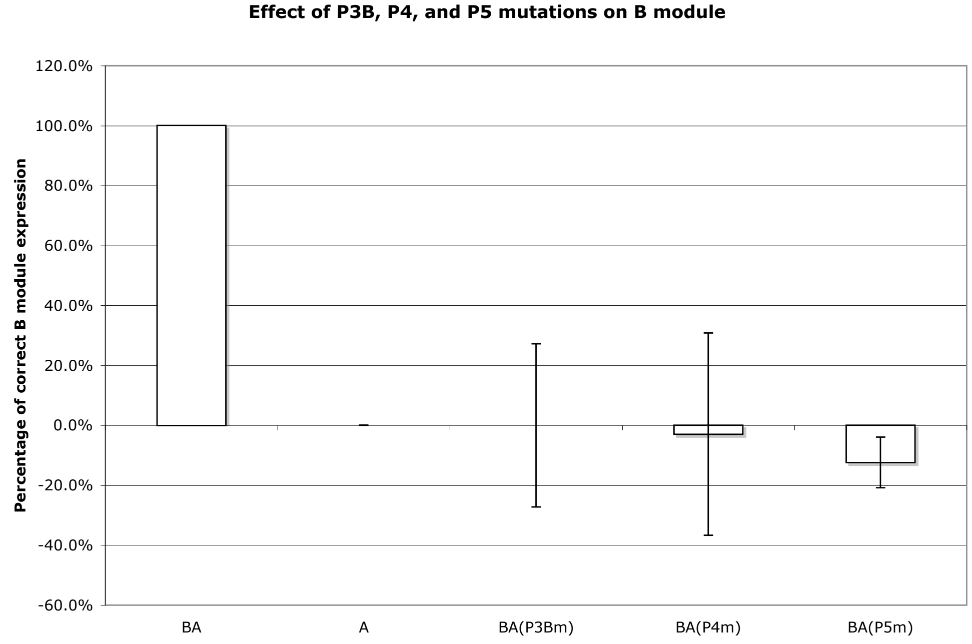


Figure 4.8: QPCR measurements of CAT reporter construct activity for BA.CAT, A.CAT, BA(P3Bm).CAT, BA(P4m).CAT, and BA(P5m).CAT at 50 hours (early gastrula). Expression levels are corrected for transgene incorporation (see Methods) and are shown as percent of correct B module expression; error bars represent standard deviation across five experiments.

no binding activity remained (data not shown). We then measured the activity of the A(P3Bm).CAT, BA(P3Bm).CAT, A(P4m).CAT, BA(P4m).CAT, A(P5m).CAT, and BA(P5m).CAT reporter constructs at 50 hrs, when module B activity is relatively high ((Kirchhamer and Davidson, 1996); Figure 4.3.1). The results (Figure 4.3.5) show unequivocally that all of the mutated constructs drive expression at the same level as module A alone. Thus each of the interactions occurring at the binding sites is individually necessary for module B function. Note that by mutating the sites we do not affect module A function at 50 hrs (Figure 4.3.4), so these three binding sites only play a role in module B at that time.

4.3.6 P3B, P4, and P5 binding sites alone are sufficient for module B function

As there are no remaining sites in *cyIIIa* with unknown function, we guessed that the P3B, P4, and P5 sites were likely to be both necessary and *sufficient* for module B function. To test this, we built BA(shortX2).CAT and A(shortX2).CAT, constructs containing two copies of an artificial version of module A bearing the three binding sites in the endogenous order. In these constructs we did not retain the spacing between sites and simply separated each site with 40 bp of randomly generated sequence (see Figure 4.3.6).

When introduced into sea urchins and measured with QPCR at 50 hrs, the BA(shortX2).CAT construct has only 10-20% more activity than the A(shortX2).CAT construct (figure 4.3.6). This additional activity must come from module B, so these sites are indeed sufficient to mediate module B activity on their own. However, we do not see the expected 2-fold increase, as with the endogenous module A (see the time course in 4.3.1, and normal BA function in figure 4.3.6). This may be because the sites were introduced without regard for inter-site spacing.

To test whether or not the precise spacing of the three binding sites matters, we built the BA(subst).GFP and A(subst).GFP constructs, which retain the P3B, P4, and P5 sites in the appropriate location and order, but with different sequence between the sites (see Methods, and figure 4.3.6). The results clearly demonstrate that the BA(subst).GFP construct exhibits a large increase over A(subst).GFP alone, indicating that the B module is indeed functional in this construct. The inclusion of the P1 sites cannot be responsible for the extra expression, because their contribution has been removed by normalizing to the A module level of expression. Thus the three sites alone are indeed sufficient to replace the communication function of module A that allows module B to drive transcription.

4.4 Discussion

4.4.1 Module A drives early expression

cyIIIa transcription starts at approximately 9 hrs (data not shown), and by 18 hrs is restricted to the presumptive aboral ectoderm. As seen in Figure 4.3.1, only module A contributes to reporter expression prior to 24 hrs.

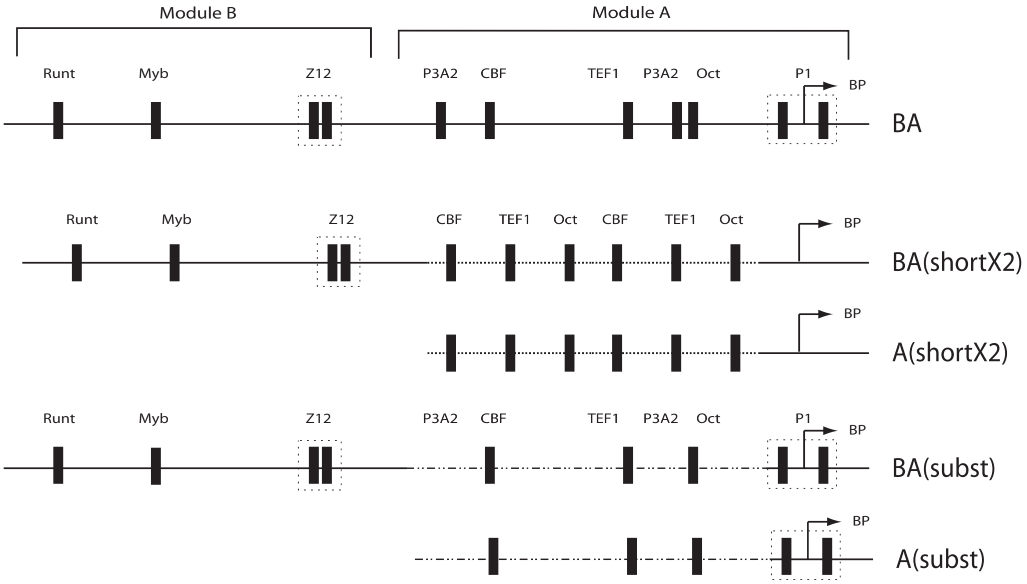


Figure 4.9: Four constructs containing the P3B, P4, and P5 sites in different arrangements. The BA(shortX2) and A(shortX2) constructs were built with an artificial A module created from three overlapping oligos containing the P3B, P4, and P5 sites in the endogenous order, equally spaced across 250 bp (see Methods). This artificial A module was then duplicated and ligated end-to-end with the B module on one side and the basal promoter on the other. The dotted sequence represents randomly generated DNA sequence. The BA(subst) and A(subst) constructs were built by embedding the P3B, P4, and P5 sites in the EpGFPII coding sequence and then fusing this sequence with the B module and the reporter construct (see Methods). The dashed sequence is the GFP coding sequence, which has no regulatory influence on its own.

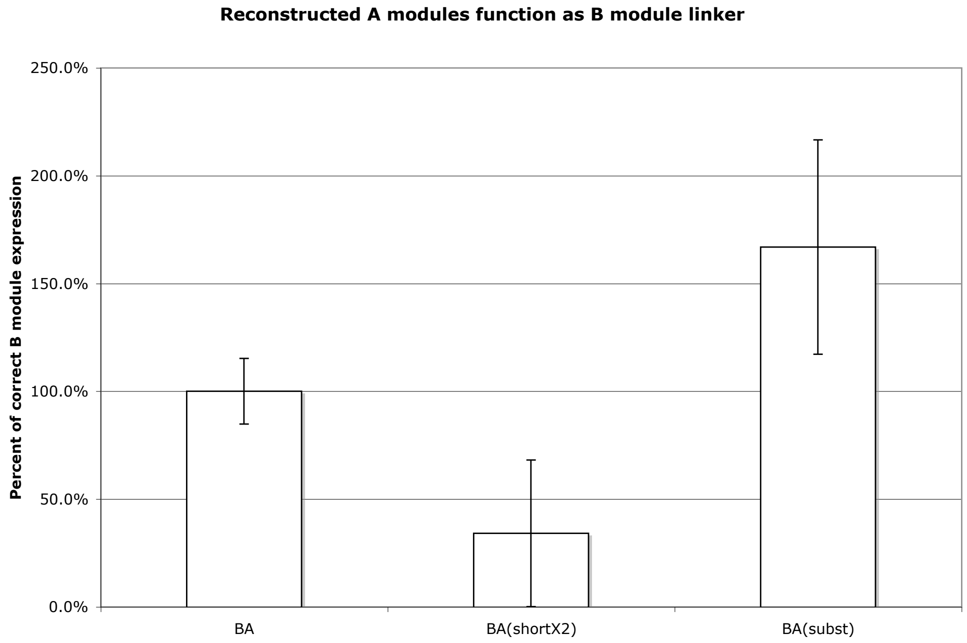


Figure 4.10: QPCR measurements of CAT or GFP expression levels of constructs bearing the wild type (BA) and artificial (BA(shortX2) and BA(subst)) sequences; see figure 4.3.6 for details. Expression levels are corrected for transgene incorporation, and are shown as percent of normal B expression levels. Measurements were taken at 50 hours. Note that in both artificial sequences (BA(shortX2) and BA(subst)) a P3A2 repressor site is eliminated, leading to stronger expression from BA(subst) than from the original wild-type sequence.

(Kirchhamer and Davidson, 1996) showed that the P1 sites in module A mediate an activation function responsible for early aboral ectoderm expression of *cyIIIa*. When either or both of the P3A2 binding sites in module A are removed, expression expands to the entire ectoderm (Kirchhamer and Davidson, 1996). Thus we know the roles of both the P1 and P3A2 binding sites.

Previous work showed that the P3B, P4, and P5 sites (respectively bound by SpOct-1, SpCBF, and SpTEF-1) were individually necessary for proper temporal expression. All three sites were shown to mediate a powerful positive function by *in vivo* titration (Franks et al., 1990). Furthermore, deletion of the P4/CBF and P5/TEF-1 sites individually caused a decrease in the number of embryos expressing CAT at levels detectable by *in situ* (Kirchhamer and Davidson, 1996).

Consonant with this result, we demonstrate that the P3B, P4, and P5 binding sites are not individually necessary for module A expression, although they do each contribute to the expression level (Figure 4.3.4). Moreover, SpOct-1, SpCBF, and SpTEF-1 are maternal factors that are probably initially present throughout the embryo, and interact with promoters that drive expression elsewhere in the embryo (Thiebaud et al., 1990). It is therefore unlikely that they direct tissue-specific expression. Thus the P1 factor is probably the only driver of module A, and it is probably responsible for the initiation of early *cyIIIa* transcription. Moreover, P1 and SpP3A2 are together responsible for the early expression of *cyIIIa* in the aboral ectoderm.

4.4.2 Module B contributes to later expression

(Coffman et al., 1996) showed that SpRunx drives module B through the P7I site. Runx-mediated activation is restricted to the aboral ectoderm by the SpZ12 and SpMyb factors, which bind at the P6 and P7II sites respectively. As there are no other positive binding sites in module B, we know that SpRunx is the only driver of the B module.

Module B drives reporter expression at approximately the same level as module A: the combined effect of both modules is roughly double the effect of A alone (Figure 4.3.1). This agrees with (Coffman et al., 2004), who measured a roughly 2-fold effect on endogenous *cyIIIa* when SpRunx was knocked down by a morpholino.

4.4.3 The P3B, P4, and P5 sites link module B with the BTA

In vivo titration and binding site deletion assays have demonstrated that the module B driver, SpRunx, can drive *cyIIIa* not only in aboral ectoderm but in oral ectoderm, endoderm, and skeletogenic mesenchyme at gastrula stages, when repressor sites are inactivated (Franks et al., 1990, Franks et al., 1990, Kirchhamer and Davidson, 1996, Coffman et al., 1997). Because the P3B, P4, and P5 sites are *necessary* for B module function, we can conclude that these factors are present and active throughout the embryo. This confirms earlier reports of the Oct-1 factor's ubiquity (Char et al., 1993, Bell et al., 1992). It also confirms that not only are CBF-A and CBF-B ubiquitous (Li et al., 2002), but that because all three subunits are necessary for CBF function, CBF-C (discussed above) must also be ubiquitously present. Finally, it suggests that TEF-1 is ubiquitously present and functional in its role as a communicator in all these tissues.

We have also shown that these three sites are both necessary and sufficient for the communicator function required for module B to drive transcription. To our knowledge, this is the first time that an entire kb of regulatory DNA has been reconstituted to precisely perform its original function by placing the known binding sites in an otherwise afunctional background sequence.

We can also conclude that DNA bending is likely to play a role in the communicator function supported by these three sites: each of these sites is separated from the other two sites by well over 100 bp, which is too far for the proteins to directly interact without bending the DNA. Moreover, when placed in close proximity the sites did not function fully (viz. BA(shortX2), Figure 4.3.6), suggesting that physical proximity on the DNA may actually detract from function.

4.4.4 A Runt-independent switch turns off the A module function of the P3B, P4, and P5 sites

At 30 hours, the P3B, P4, and P5 sites in module A contribute to module A output, while at 50 hours, they no longer contribute (Figure 4.3.4). This behavior is reminiscent of the UI concentration-dependent switch in *endo16 cis*-regulatory region (Yuh et al., 2001). However, unlike in *endo16*, this switch does not require the presence of module B: interactions at the P3B, P4, and P5 sites no longer contribute to module A at 50 hours *whether or not* module B is present. Nor is the switch away from module A amplification mediated by any of the three factors: removing an individual binding site does not restore the module A function of the other two at 50 hours, for any of the three sites. This suggests that the switch in function is either mediated by P1 or is regulated externally to the factors binding directly to the *cis*-regulatory region.

4.4.5 Modules A and B are logically distinct

The drivers for modules A and B are different sets of proteins that act at distinct times and are present in different territories. Module A responds to an early ectoderm-specific input and is repressed on the oral side by SpP3A2, while module B is driven by the ubiquitous Runx factor and repressed specifically by two proteins, SpZ12 and SpMyb. Module A drives *cyIIIa* from 9 hrs onward; module B only starts driving *cyIIIa* transcription after 18 hrs. Thus while these modules drive expression in the same territory at overlapping times, they do so in response to different drivers and repressors.

The spatial repression functions are combined only after interactions internal to the modules take place: for example, removing the SpZ12 binding sites does not lead to ectopic expression by module A. The functions are logically distinct, in the sense that either module works independently of the drivers of the other module, and their output is summed.

Module C, however, operates *after* the outputs of module A and B are produced, and acts on the combined output: even if either module's driver function is abrogated, module C amplifies the output of the remaining module (Kirchhamer and Davidson, 1996, Coffman et al., 1997).

These three modules fulfill three different roles: module A responds to P1, an early specification-level driver. Module B responds to SpRunx, a factor that is probably a general driver of cell proliferation and differentiation (Robertson et al., 2002). Module C acts as a hardwired amplifier, adjusting the amplitude of the output without affecting its location or timing. This modularity of function seems to be a general feature of *cis*-regulatory regions, perhaps because it leads to *separability* of function, which could increase both robustness and evolvability of *cis*-regulatory regions.

4.4.6 A logic model of the cyIIIa *cis*-regulatory region

In Figure 4.4.6, we present a logic diagram representing the information processing performed by the *cyIIIa cis*-regulatory region. This "view from the genome" links the presence or absence of a specific *cis*-regulatory interaction to the ultimate regulatory yield of the entire *cis*-regulatory region. And, as with *endo16*, each interaction yields a number of additional testable hypotheses.

The *cyIIIa cis*-regulatory region adds several examples to our *cis*-regulatory lexicon. There are four distinct types of logic interactions present in *cyIIIa*:

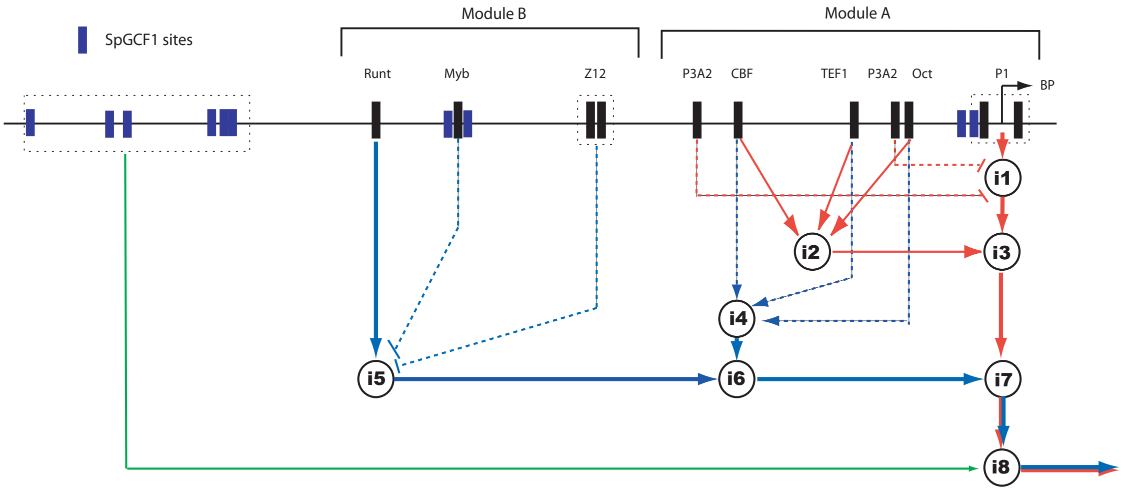
1. The activation ("driver") functions driven by P1 in module A and Runx in module B.

2. The combinatorial logic operators responsible for repressing driver activation within each module, mediated by SpP3A2 in module A and SpMyb and SpZ12 in module B.

3. The looping/linker function fulfilled by the many SpGCF1 binding sites throughout the *cis*-regulatory region, which may function biochemically to bring both repressors and activators into physical contact with the basal transcription apparatus.

4. The newly discovered "communicator" function mediated by the P3B/Oct-1, P4/CBF, and P5/TEF-1 binding sites, which is required for module B function. In *endo16*, the R linker site acts as a switch and transfers control from module A to module B when the Otx protein is present in sufficient concentration. In *cyIIIa*, however, the linking factors are present in early embryogenesis even when the SpRunx activating factor is not, which suggests that the link is constitutively active and module B output is entirely driven by the SpRunx factor.

The *cyIIIa cis*-regulatory region also contains sites that play different roles in early and late function. Measured at 30 hours, the presence of the P3B, P4, and P5 sites increases the expression driven by module A (Figure 4.3.4). By 50 hours, however, they no longer participate in module A expression. At 50 hours, these same three sites are required for the activity of module B, indicating that they act in concert with SpRunx to drive expression. How this switch in function occurs is unclear, but it cannot be driven solely by SpRunx, as the experiments with module A alone lack a SpRunx site.



if P1 and not P3A2	i1 = P1(t)
else	i1 = 0
	i2 = (Oct + CBF + TEF)
if t < t0	i3 = i1 * (1 + i2)
else	i3 = i1
if (Oct and CBF and TEF-1)	i4 = 1
else	i4 = 0
if Runt and not (Myb or Z12)	i5 = Runt(t)
else	i5 = 0
if i4	i6 = i5
else	i6 = 0
i7 = i1 + i6	
if GCF-1	i8 = 2 × i5
else	i8 = i5

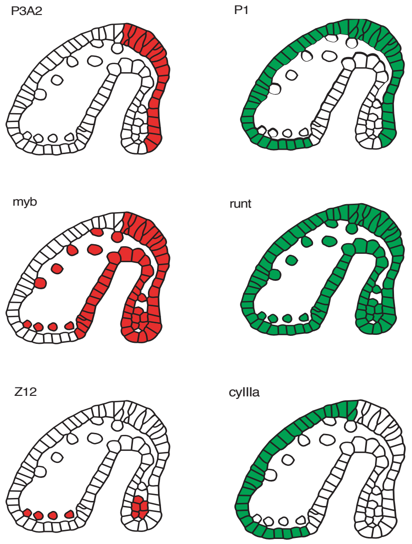


Figure 4.11: A computational logic model for the entire *cyIIIa cis*-regulatory system. The regulatory DNA of *cyIIIa* is shown as a horizontal strip at the top of the diagram. The individual binding sites are indicated by labeled boxes. Module C amplification is shown in green; module B and its effects are shown in blue; module A and its effects are shown in red. Intermediate logic functions (i) are indicated by numbered circles. Each represents a specific regulatory interaction modeled as a logic operation. The role of each regulatory interaction in the overall regulation of *cyIIIa* is indicated by the lines emanating from each binding site: as in the *endo16* model of (Yuh et al., 2001), thick solid lines represent time-varying driver functions; thin solid lines represent amplifying functions; and dashed lines represent Boolean operations (repression or communication). The logic functions that represent logical interactions between transcription factors are defined in the statements below the diagram. Note in particular the time dependence of i3 that determines whether or not the Oct, CBF, and TEF factors contribute to module A function (figure 4.3.4; see text).

We may also start to generalize about repressor function. In *cyIIIa*, repression appears to be local to both module A and B. In *endo16*, spatial repression is performed by the DC, E, and F modules and mediated through the R site in module A. Despite this difference, in both *endo16* and *cyIIIa* the repressors act as absolute negatives, repressing transcription entirely in tissues in which they are present. It is likely that this sharp switch-like response is a general logical feature of repressors.

4.4.7 Conclusions

cis-Regulatory regions perform a important role as information processors, integrating spatial and temporal cues to produce specific expression patterns. This role underscores the importance of understanding the precise functions encoded by *cis*-regulatory regions. Ultimately we would like to decode these regions computationally; however, we cannot analyze the functions encoded by *cis*-regulatory regions computationally without first understanding many more *cis*-regulatory regions experimentally (Istrail and Davidson, 2005). The addition of cyIIIa to the lexicon of logically understood *cis*-regulatory regions doubles our previous state of knowledge.

Bibliography

Angerer, R. and Davidson, E., 1984. Molecular indices of cell lineage specification in sea urchin embryos. *Science*, **226**(4679):1153–60.

Barberis, A., Superti-Furga, G., and Busslinger, M., 1987. Mutually exclusive interaction of the ccaat-binding factor and of a displacement protein with overlapping sequences of a histone gene promoter. *Cell*, **50**(3):347–59.

Bell, J., Char, B., and Maxson, R., 1992. An octamer element is required for the expression of the alpha h2b histone gene during the early development of the sea urchin. *Dev Biol*, **150**(2):363–71.

Bogarad, L., Arnone, M., Chang, C., and Davidson, E., 1998. Interference with gene regulation in living sea urchin embryos: transcription factor knock out (tko), a genetically controlled vector for blockade of specific transcription factors. *Proc Natl Acad Sci U S A*, **95**(25):14827–32.

Calzone, F., Hoog, C., Teplow, D., Cutting, A., Zeller, R., Britten, R., and Davidson, E., 1991. Gene regulatory factors of the sea urchin embryo. i. purification by affinity chromatography and cloning of p3a2, a novel dna-binding protein. *Development*, **112**(1):335–50.

Calzone, F., Theze, N., Thiebaud, P., Hill, R., Britten, R., and Davidson, E., 1988. Developmental appearance of factors that bind specifically to cis-regulatory sequences of a gene expressed in the sea urchin embryo. *Genes Dev*, **2**(9):1074–88.

Campbell, S., Inamdar, M., Rodrigues, V., Raghavan, V., Palazzolo, M., and Chovnick, A., 1992. The scalloped gene encodes a novel, evolutionarily conserved transcription factor required for sensory organ differentiation in drosophila. *Genes Dev*, **6**(3):367–79.

Char, B., Bell, J., Dovala, J., Coffman, J., Harrington, M., Becerra, J., Davidson, E., Calzone, F., and Maxson, R., 1993. Spoct, a gene encoding the major octamer-binding protein in sea urchin embryos: expression profile, evolutionary relationships, and dna binding of expressed protein. *Dev Biol*, **158**(2):350–63.

Coffman, J. and Davidson, E., 2001. Oral-aboral axis specification in the sea urchin embryo. i. axis entrainment by respiratory asymmetry. *Dev Biol*, **230**(1):18–28.

Coffman, J., Dickey-Sims, C., Haug, J., McCarthy, J., and Robertson, A., 2004. Evaluation of developmental phenotypes produced by morpholino antisense targeting of a sea urchin runx gene. *BMC Biol*, **2**:6.

Coffman, J., Kirchhamer, C., Harrington, M., and Davidson, E., 1996. Sprunt-1, a new member of the runt domain family of transcription factors, is a positive regulator of the aboral ectoderm-specific cyiiia gene in sea urchin embryos. *Dev Biol*, **174**(1):43–54.

Coffman, J., Kirchhamer, C., Harrington, M., and Davidson, E., 1997. Spmyb functions as an intramodular repressor to regulate spatial expression of cyiiia in sea urchin embryos. *Development*, **124**(23):4717–27.

Cox, K., Angerer, L., Lee, J., Davidson, E., and Angerer, R., 1986. Cell lineage-specific programs of expression of multiple actin genes during sea urchin embryogenesis. *J Mol Biol*, **188**(2):159–72.

Flytzanis, C., Britten, R., and Davidson, E., 1987. Ontogenic activation of a fusion gene introduced into sea urchin eggs. *Proc Natl Acad Sci U S A*, **84**(1):151–5.

Franks, R., Anderson, R., Moore, J., Hough-Evans, B., Britten, R., and Davidson, E., 1990. Competitive titration in living sea urchin embryos of regulatory factors required for expression of the cyiiia actin gene. *Development*, **110**(1):31–40.

Hoog, C., Calzone, F., Cutting, A., Britten, R., and Davidson, E., 1991. Gene regulatory factors of the sea urchin embryo. ii. two dissimilar proteins, p3a1 and p3a2, bind to the same target sites that are required for early territorial gene expression. *Development*, **112**(1):351–64.

Hough-Evans, B., Britten, R., and Davidson, E., 1988. Mosaic incorporation and regulated expression of an exogenous gene in the sea urchin embryo. *Dev Biol*, **129**(1):198–208.

Hough-Evans, B., Franks, R., Zeller, R., Britten, R., and Davidson, E., 1990. Negative spatial regulation of the lineage specific cyiiia actin gene in the sea urchin embryo. *Development*, **110**(1):41–50.

Istrail, S. and Davidson, E., 2005. Logic functions of the genomic cis-regulatory code. *Proc Natl Acad Sci U S A*, **102**(14):4954–9.

Jacquemin, P., Hwang, J., Martial, J., Dolle, P., and Davidson, I., 1996. A novel family of developmentally regulated mammalian transcription factors containing the tea/atts dna binding domain. *J Biol Chem*, **271**(36):21775–85.

Kirchhamer, C. and Davidson, E., 1996. Spatial and temporal information processing in the sea urchin embryo: modular and intramodular organization of the cyiiia gene cis-regulatory system. *Development*, **122**(1):333–48.

Lee, J., Calzone, F., and Davidson, E., 1992. Modulation of sea urchin actin mrna prevalence during embryogenesis: nuclear synthesis and decay rate measurements of transcripts from five different genes. *Dev Biol*, **149**(2):415–31.

Li, X., Bhattacharya, C., Dayal, S., Maity, S., and Klein, W., 2002. Ectoderm gene activation in sea urchin embryos mediated by the ccaat-binding factor. *Differentiation*, **70**(2-3):109–19.

Li, X., Mantovani, R., van Huijsduijnen, R. H., Andre, I., Benoist, C., and Mathis, D., 1992. Evolutionary variation of the ccaat-binding transcription factor nf-y. *Nucleic Acids Res*, **20**(5):1087–91.

Ransick, A. and Davidson, E., 2006. cis-regulatory processing of notch signaling input to the sea urchin glial cells missing gene during mesoderm specification. *Dev Biol*, **297**(2):587–602.

Revilla i Domingo, R., Minokawa, T., and Davidson, E., 2004. R11: a cis-regulatory node of the sea urchin embryo gene network that controls early expression of spdelta in micromeres. *Dev Biol*, **274**(2):438–51.

Robertson, A., Dickey, C., McCarthy, J., and Coffman, J., 2002. The expression of sprunt during sea urchin embryogenesis. *Mech Dev*, **117**(1-2):327–30.

Shott, R., Lee, J., Britten, R., and Davidson, E., 1984. Differential expression of the actin gene family of strongylocentrotus purpuratus. *Dev Biol*, **101**(2):295–306.

Thiebaud, P., Goodstein, M., Calzone, F., Theze, N., Britten, R., and Davidson, E., 1990. Intersecting batteries of differentially expressed genes in the early sea urchin embryo. *Genes Dev*, **4**(11):1999–2010.

Vaudin, P., Delanoue, R., Davidson, I., Silber, J., and Zider, A., 1999. Tondu (tdu), a novel human protein related to the product of vestigial (vg) gene of drosophila melanogaster interacts with vertebrate tef factors and substitutes for vg function in wing formation. *Development*, **126**(21):4807–16.

Wang, D., Britten, R., and Davidson, E., 1995a. Maternal and embryonic provenance of a sea urchin embryo transcription factor, spz12-1. *Mol Mar Biol Biotechnol*, **4**(2):148–53.

Wang, D., Kirchhamer, C., Britten, R., and Davidson, E., 1995b. Spz12-1, a negative regulator required for spatial control of the territory-specific cyiiia gene in the sea urchin embryo. *Development*, **121**(4):1111–22.

Yuh, C., Bolouri, H., and Davidson, E., 2001. Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development*, **128**(5):617–29.

Yuh, C., Dorman, E., Howard, M., and Davidson, E., 2004. An otx cis-regulatory module: a key node in the sea urchin endomesoderm gene regulatory network. *Dev Biol*, **269**(2):536–51.

Zeller, R., Britten, R., and Davidson, E., 1995a. Developmental utilization of spp3a1 and spp3a2: two proteins which recognize the same dna target site in several sea urchin gene regulatory regions. *Dev Biol*, **170**(1):75–82.

Zeller, R., Coffman, J., Harrington, M., Britten, R., and Davidson, E., 1995b. Spgcf1, a sea urchin embryo dna-binding protein, exists as five nested variants encoded by a single mrna. *Dev Biol*, **169**(2):713–27.

Zeller, R., Griffith, J., Moore, J., Kirchhamer, C., Britten, R., and Davidson, E., 1995c. A multimerizing transcription factor of sea urchin embryos capable of looping dna. *Proc Natl Acad Sci U S A*, **92**(7):2989–93.

Chapter 5

Evolutionary Comparisons Suggest Many Novel cAMP Response Protein Binding Sites in *E. coli*

This work is already published.[1]

5.1 Introduction

The binding of transcription factors (TFs) to specific sites is a central mechanism of transcriptional regulation (1). Powerful computational techniques for finding putative binding sites in genomes and for characterizing transcription factor binding on whole-genome scales are becoming available (2, 3, 4, 5, 6). An energy matrix of size $4 \times L$ (where L is the site length) is often used to capture the binding profile of a particular transcription factor, under the simplifying assumption that the contribution of a particular position in a binding site is independent of neighboring positions (7, 8, 9). This assumption is of unknown accuracy and, even if it is a good physical approximation, the particular choice of energy matrix could be inadequate, leading to false identification of sites.

The cAMP Response Protein (CRP) is a dimeric DNA-binding and DNA-bending protein that binds in multiple *E. coli* promoters to 22 bp sites with a core consensus sequence `GTGANNNNNNTCAC` (10); the DPInteract database (11) contains 48 such sites, while RegulonDB (12) contains more than 88. Computational studies typically predict many more CRP binding sites than are found in either of these databases (13, 14, 15). While the databases are certainly not complete, the apparent large overprediction of binding sites undermines the credibility of the matrix model of binding specificity on which these studies are based. To sharpen the issue, we contrast the situation for CRP with that for LacI, a TF which has a small number of known binding sites and is thought to be highly specific. In fact, an energy matrix which cleanly discriminates the known binding sites from the rest of the genome can be found for LacI. To decide whether CRP binding can successfully be described by a similar matrix model, we need to know whether the overpredicted CRP binding sites are functional. We look for evidence on this issue by identifying partner sites in orthologous regions in *S. typhimurium*, a close relative of *E. coli*, and examining mutations between site pairs. The mutation probability depends on position in the site in a way that ensures rough conservation of CRP binding energy between species. We take this as evidence that the novel CRP sites have real biological function. We do not know why experimental methods have not picked them up, but the evidence that evolution cares about them is striking. We conclude that the simple matrix model works as well for CRP as for LacI, despite their very different binding profiles. We also make specific predictions of many new CRP binding sites (and regulated genes).

5.2 Methods

Genomes. We work with the genomes of *E. coli K12* (NCBI accession number NC 004431, 4,639,221 bp) and *S. typhimurium LT2* (NCBI accession number NC 003197, 4,857,432 bp). Genes and intergenic regions

[1] Brown CT, Callan CG Jr. *Evolutionary comparisons suggest many novel cAMP response protein binding sites in Escherichia coli.* **Proc Natl Acad Sci U S A.** 2004 Feb 24;101(8):2404-9.

are identified by comparison with protein tables available from NCBI. Genomes and tables are included in the Software distribution (below).

Matrix construction. A given TF contacts the genome at a site of length L (of order 20 base pairs in typical bacterial examples). To estimate the sequence–dependent affinity, we assign to each TF a matrix $\sigma_{b,i}$ which is used to score a site $(b(i), i = 1 \ldots L)$ according to the additive rule

$$E = \sum_{i=1}^{L} \sigma_{b(i),i} \ .$$

We will refer to σ as the energy matrix of the TF and E is meant to approximate the binding energy (in units of kT) of the TF to the site. The usual method for constructing σ (7) starts from a list of known binding sites for a TF and applies the following algorithm: for each position i in a site, the number of occurrences $N_i(b)$ of each DNA base b in the list of sites is counted, and the matrix elements are assigned by the rule

$$\sigma_{bi} = \log \frac{\max_a N_i(a) + 1}{N_i(b) + 1} \ .$$

(see Berg & von Hippel, (16)). The +1 pseudocount regularizes the divergence that arises if any of the $N_i(a)$ happen to vanish in the finite sample of known sites. The matrix is normalized to assign $\sigma_{b,i} = 0$ to the most common base pair at site i and $\sigma_{bi} > 0$ to all others; thus, the consensus sequence has $E = 0$, and all others have $E > 0$. For an implementation of this algorithm, see the `openfill` program distributed with the software (see Software below). This method for assigning values to σ_{bi} has a sound physical and evolutionary rationale (7) when all the input site sequences have roughly the same binding energy to the TF. When this is not the case, a more sophisticated algorithm may be needed (a concrete example will be given shortly). In short, the validity of the matrix model is distinct from the validity of the algorithm for evaluating the matrix itself.

The work reported here specifically concerns two TFs: CRP and LacI. We used the known sites in the *E. coli K12* genome as listed in DPInteract (11). LacI contacts 21 bp and has three listed binding sites (all in close proximity to the *lac* operon); CRP contacts 22 bp and has 48 listed binding sites (widely dispersed in the genome). Binding site files suitable for use with the `openfill` program are available (see Software below). These files include twice as many sequences as cited above (i.e 6 for LacI and 96 for CRP). This is because CRP and LacI, like many bacterial TFs, are symmetric dimers and can be regarded as reading either the top or the bottom strand of the DNA. The two strand reads are not usually identical and can legitimately be cited as independent site data. However, as is appropriate for CRP and LacI, `openfill` creates a symmetrized energy matrix which assigns the same energy to a sequence and its reverse complement (and therefore to both reads of a site).

Binding site search. Binding site search was done with a program `scangen` (see Software, below) which takes as input the energy matrix for a transcription factor, a genome and a file containing the bounding coordinates and names of all coding regions in that genome (the latter two obtained from the NCBI database as described above). The scan program assigns an energy score to each site in the genome and produces two types of output: cumulative histograms of number of binding sites versus E-value (both total number of sites and the number of sites in non-coding regions) and a list of all sites below a chosen cutoff E-value, giving for each site its coordinate and, if located in a non-coding region, the names of the flanking genes. The data concerning location with respect to coding regions is a useful diagnostic because functional sites are mostly located in non-coding regions.

Orthology and alignment of intergenic regions and site pairs. We declare two intergenic regions to be orthologous if the flanking genes in both *E. coli* and *S. typhimurium* have the same gene names (according to the NCBI protein tables) and if the regions align well with CLUSTALW (discussed below). The two genomes have extensive orthology: there are 3475 intergenic regions in *E. coli* and 3660 in *S. typhimurium*, of which 1533 are orthologous by this definition. This is a rather restrictive notion of orthology; other workers (6, 17) find an additional ~ 500 orthologous intergenic regions based on orthology of downstream proteins. With either definition, the mean difference rate (disregarding end gaps) between orthologous intergenic region pairs is comparable (about 25%).

For an *E. coli* site lying in an intergenic region having an *S. typhimurium* ortholog, we identify the sequence of the orthologous site by alignment of the relevant pair of intergenic regions. We used CLUSTALW v1.83 (18) to align intergenic region pairs, keeping 30bp of coding sequence on either side (typically a few hundred bp of sequence in all); the default parameters for CLUSTALW were used. The *S. typhimurium* sequence aligned to the *E. coli* site is then defined to be its orthologous partner site, provided the alignment places no gaps in either sequence. We place no limitation on the number of mutations between the two sites.

Calculation of expected energy change under mutation. We calculate the expected energy change for a site by first calculating the average rate of transitions and transversions for each region in *E. coli* from the CLUSTALW alignment between the region and its orthologous region in *S. typhimurium.* We adjust the rate of transitions by a factor of two to account for null- and back-mutations. Then, for each purine in the site we calculate the expected energy change to be

$$\delta\, e(i) = p_{\mathrm{transition}} < e_{\mathrm{purine}}(i) > + p_{\mathrm{transversion}} < e_{\mathrm{pyrimidine(i)}} >$$

where $< e_{\mathrm{purine}} >$ and $< e_{\mathrm{pyrimidine}} >$ are the average contributions of a purine and a pyrimidine at that position in the matrix. The contribution of each pyrimidine is calculated similarly, and the $\delta\, e$ values are summed over all positions to calculate the total expected change in energy. We did not introduce gapping into our model because gaps in sites in either genome disqualify a site for comparison and, moreover, occur

Table 1: Sites isolated from the genome at several energy cutoffs with `LacI-naive`, a matrix representing LacI binding. The matrix was constructed from the three binding sites implicated in transcriptional regulation of the *lac* operon. Mean (standard deviation) of energy over all sites in the genome is 20.9 (3.2).

Cutoff	**# sites**	**coding**	**intergenic**	**known**
3.00	1	0%	100%	1/3
5.00	2	50%	50%	2/3
7.00	26	85%	15%	3/3
9.00	483	86%	14%	3/3

infrequently (less than 5% of the aligned sequence consists of gaps).

Software and Software Availability. All software was developed in C++ and/or Python 2.3 under Linux. The software was developed *ab initio* by the authors; a package that can be used to reproduce all of our results is available at `http://www.princeton.edu/~ccallan/binding/` The software is Copyright (C) Princeton University and the California Institute of Technology, 2003; it is freely available and redistributable under an Open Source-compatible license.

5.3 Results

Binding profiles of CRP and LacI differ. LacI and CRP have very different profiles: binding site catalogs indicate that LacI has three known sites in the immediate vicinity of the *lac* operon, while CRP affects the transcription of many genes and has 48 listed sites. While it may be too strong to say that LacI regulates only the *lac* operon, it does seem clear that CRP affects many more genes than LacI. Since it is not obvious that the linear additive model for binding affinity can encompass both extreme behaviors, a comparative study of these two cases should be instructive.

We created energy matrices `LacI-naive` and `CRP-naive` using the known sites as described in Methods and then used `scangen` to create histograms of the energy distribution of all sites in the *E. coli* genome. The results are presented in Tables 1 and 2. Each line gives: the cumulative number of sites with E less than the indicated cutoff, the cumulative percentage of sites that are in genes, the cumulative percentage in intergenic regions and the cumulative number of the known (or input) sites that have been captured. We count sites in coding and non-coding regions separately because functional sites are expected to lie mostly in the non-coding 15% of the *E. coli* genome and the statistics of site location could convey useful evidence for or against functionality.

The tables display a potential problem with the linear additive model: using E-value as the discriminant, the model seems to over-predict the number of TF binding sites. For LacI, note that while the top two bins in Table 1 capture only known sites, the next bin captures the remaining known site plus 23 others. The

Table 2: Sites isolated from the genome at several energy cutoffs with `CRP-naive`, a matrix constructed from the list of 48 known CRP binding sites. Mean (standard deviation) of energy over all sites in the genome is 26.9 (4.8).

Cutoff	**# sites**	**coding**	**intergenic**	**known**
5.00	31	3%	97%	4/48
7.00	105	9%	91%	10/48
9.00	375	26%	74%	27/48
11.00	1495	53%	47%	39/48
15.00	26873	72%	28%	48/48

novel sites, while competitive in E-value, are probably not true LacI sites since they are randomly located in the genome, with no preference for non-coding regions. Even so, the `LacI-naive` matrix discriminates the known sites from the rest of the genome surprisingly well, given how little input data is used. The situation for CRP, displayed in Table 2, is less good: while the known sites are assigned small E-values compared to a random site in the genome, they have a large range in E ($\Delta E \sim 12$, or five orders of magnitude in calculated affinity) and many sites not in the known site list have comparable or better E-values. Since CRP regulates many genes, it is less clear than for LacI that the novel sites are spurious. In contrast to LacI, the novel sites are not randomly distributed: the lower the E-value, the more likely they are to lie in non-coding regions. This suggests that many of the novel CRP sites may be real. We will develop other lines of evidence for this in what follows.

A modified LacI matrix accurately discriminates known sites. It is important to understand whether the above problems are due to a non-optimal choice of the matrix or instead to a failure of the linear additive model for sequence-dependent binding affinity. To explore this, we also ran `scangen` on two other matrices, `LacI-relax` and `CRP-relax`, both derived using a relaxation method to be described elsewhere (19). The relaxation method takes as input the known sites and their relative binding affinities (if known) as well as the background genome. The use of the relative binding affinities is crucial: it has the effect that the sequences of subsidiary weak binding sites, should any exist, can be used as input data to refine the matrix without destabilizing it. The relaxation implements the notion that for proper function, not only must binding to the known sites be strong, but net binding to the rest of the genome must be small. The matrices produced by this method differ significantly from the "naive" matrices and lead to the site histograms displayed in Tables 3 and 4 (the matrices used are available on-line as discussed in Software).

Comparing Table 3 with Table 1, we see that `LacI-relax`, in contrast with `LacI-naive`, succeeds in creating a gap in energy between the known sites and the rest of the genome, thus realizing the picture that LacI acts primarily on the *lac* operon. This is not a prediction, just the (non-trivial) statement that a linear model matrix that fits the qualitative facts about LacI can be found. Whether this matrix correctly predicts

Table 3: Site list for `LacI-relax`, a matrix constructed to pick out known LacI sites preferentially. Note that two of the three known LacI binding sites are located in coding regions.

Cutoff	**# sites**	**coding**	**intergenic**	**known**
3.00	1	0%	100%	1/3
5.00	1	0%	100%	1/3
7.00	3	33%	67%	3/3
9.00	7	71%	29%	3/3

Table 4: Site list for `CRP-relax`, a matrix constructed to pick out known CRP sites preferentially.

Cutoff	**# sites**	**coding**	**intergenic**	**known**
5.00	103	10%	90%	11/48
7.00	524	38%	62%	33/48
9.00	3545	66%	34%	43/48
11.00	21386	74%	26%	48/48
15.00	275384	79%	21%	48/48

finer details like the affinity distribution of weaker (presumably non-functional) sites below the gap is an interesting question (to be explored elsewhere). If other genes than the *lac* operon were to be found to be regulated by LacI, this discussion would have to be revisited.

Table 4 shows the result of an attempt to improve the matrix for CRP by a similar relaxation method: apart from a possible rescaling of the E-value, Table 4 looks similar to Table 2. In both tables, the known sites span a wide range of E-values and are accompanied by a large number of novel sites of comparable E-values. In what follows, we will look for evidence concerning the functionality of these sites via interspecies comparisons. This is important since, if the novel sites are spurious, it will be hard to avoid concluding that the linear additive energy model fails as a framework for predicting functionality, at least for CRP. Since relaxation did not have much effect, we will revert to assessing CRP sites using the matrix `CRP-naive`.

Novel CRP sites have similar binding energies in *S. typhimurium.* To assess the status of the novel *E. coli* sites found by the CRP energy matrix, we performed comparisons with orthologous sites in the closely-related organism *S. typhimurium.* Partners of strong *E. coli* sites lying in intergenic regions are constructed by alignment of the orthologous *S. typhimurium* intergenic region (see Methods). Since the CRP energy matrix is not used in generating the alignment, we have no reason to expect the aligned sites in *S. typhimurium* to be scored as strong binding sites. (Note that *S. typhimurium* CRP differs from *E. coli* CRP in only one of 210 amino acids, and can be expected to have the same sequence-specific binding affinity.)

What actually happens is shown in Figure 1 where we compare the energy of strong CRP sites in *E. coli* with the energy of orthologous sites in *S. typhimurium.* The graph shows the line of equality (no mutations),

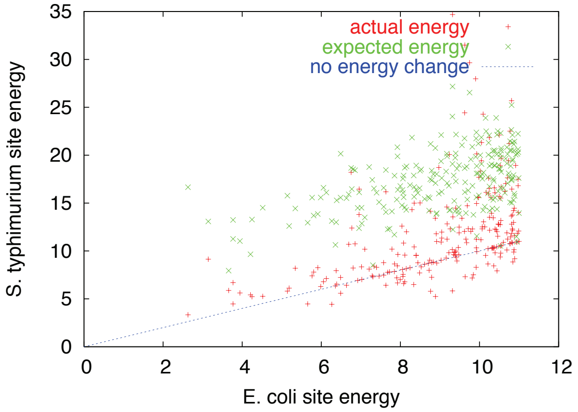


Figure 5.1: Binding energy is strongly correlated between orthologous site pairs.

a scatter plot of the actual orthologous pair energies (actual mutations), and a scatter plot of the same sites with the *S. typhimurium* site energy replaced by its expected value under random mutations. Actual mutated energies lie well below what is expected on the random mutation model for *E. coli* site energies up to $E \sim 7$ (and a population difference exists up to higher values of site energy). The fixation of energy is not due to a general suppression of mutation: for the 34 site pairs with *E. coli* site energy less than 7.0, the average number of mutations per site is 4.1, not very different from the background rate of 5.5 for 22 bp sequences. The actual number of mutations per site runs from 0 to 11 and highly mutated sequences often have nearly the same energy (in one case, 7 out of 22 positions are mutated, yet the change in E is only .22). The conservation of CRP binding energy (as distinct from sequence conservation) suggests that these sites are functional for CRP in some way useful to the organism.

Novel CRP site pairs have strongly biased mutational patterns. The most striking evidence of conservation is obtained by computing population averages of the position dependence of the probability of mutation (20). We did this for populations of orthologous site pairs defined by several energy cutoffs. The results, displayed in Figure 2, show a strong positional bias within the site (the data have been reflected about the center position to smooth out statistical noise). The bias washes out abruptly when the cutoff exceeds $E_{\text{cut}} \simeq 11$. This is beyond the energy where, according to Figure 1, the close correlation between individual site energies in the two species begins to wash out.

For all cutoffs, novel sites greatly outnumber input sites. Thus, the positional bias in mutation frequency is a property of the novel sites. The same analysis for the input sites alone (38 of the 48 input sites have *S. typhimurium* orthologs) gives a result indistinguishable from the $E_{\text{cut}} = 9$ curve in Figure 2 (see Supporting Information). The positional bias in mutation frequency is essentially the same for the novel sites and the input sites and presumably has a common cause.

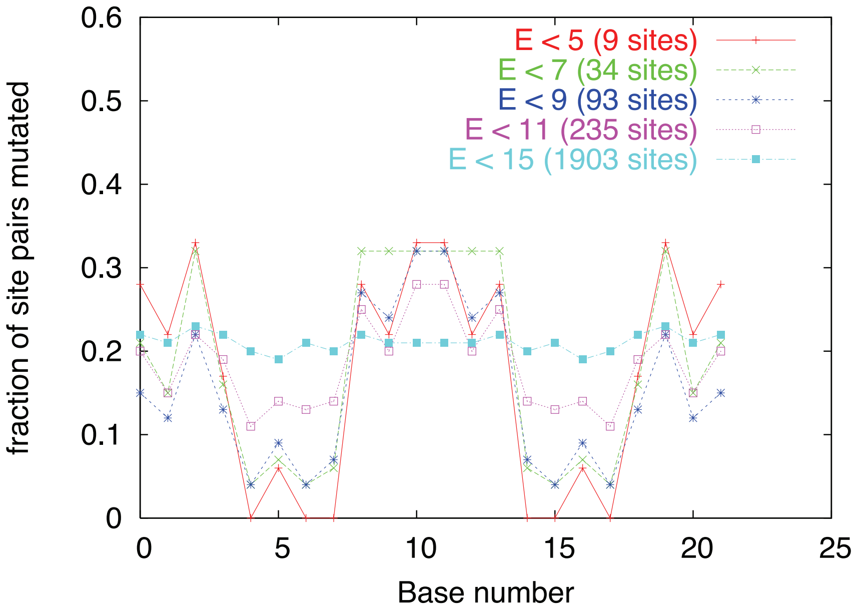


Figure 5.2: Paired site populations show a position-dependent pattern of mutation rates.

The positional bias has the shape expected if the sites are constrained to strongly bind a dimeric transcription factor in both species. For example, only two of the site pairs found with $E_{\text{cut}} = 5$ have even a single mutation in the core positions 4-7 and their complements at positions 14-17. This pattern of fixation suggests that these eight bases dominate the site energy, consistent with the usual picture of a core region of the transcription factor contacting the DNA and selecting a central motif. An examination of the `CRP-naive` matrix does show that these positions should be the most strongly fixed. We emphasize that the *S. typhimurium* sites were chosen with no consideration of their CRP binding energy: the non-random pattern of the mutations between them and their *E. coli* orthologs is evidence that some biologically meaningful aspect of the site sequence is being conserved. A plausible interpretation is that the sites are functional and that their CRP binding energy is approximately conserved.

Figure 2 also confirms that the correlation of binding energies between *E. coli* and *S. typhimurium* shown in Figure 1 is not due to strict sequence conservation: in the population of sites with $E_{\text{cut}} = 7$, only 8 of the 22 bases have a mutation rate of less than 5%, while the 14 remaining bases vary at a rate of 18% or higher. The total number of mutations between orthologous site sequences is often large and it would be difficult to identify the orthologous pairs by looking for local patches of stronger-than-background sequence conservation.

Table 5: Scangen list for CRP binding, using a matrix modified to ignore contributions from any but the 8 core base pairs of the site (see Figure 2). The list of known sites in this table is chosen as all those sites that score with an energy of 9 or less using `CRP-naive`.

Cutoff	**# sites**	**coding**	**intergenic**	**known**
1.00	166	51%	49%	70/375
3.00	2508	82%	18%	217/375
5.00	17411	86%	14%	325/375
7.00	78995	86%	14%	372/375
9.00	271661	86%	14%	375/375

Flanking bases are required for discrimination ability. The pattern of base fixation shown in Figure 2 suggests that there is a strong constraint on 8 core positions within each site. This raises the question whether the flanking sequence lends useful discriminatory ability. To study this, we constructed a `CRP-core-only` matrix from `CRP-naive` by setting to zero all the entries in the rows corresponding to the 14 flanking positions. The site energy histogram produced by scanning the *E. coli* genome with this modified matrix is given in Table 5. The high-scoring sites clearly constitute a very different population from the high-scoring sites under the `CRP-naive` matrix. There are many ways to see this, but an examination of the sites in the first bin ($E < 1$) makes the point: these sites all have $E \equiv 0$, *i.e.* they match the consensus sequence for the 8 core positions. But half of these "perfect" sites lie in coding regions, while the sites that our evolutionary analysis suggests are under control of CRP are very heavily concentrated in the non-coding regions. The situation gets progressively worse for higher cutoffs. We conclude that the core-only matrix cannot sharply discriminate the sites of interest. Individual flanking positions contribute limited information but, taken together, they dramatically enhance the ability of the matrix to discriminate functional sites (or at least the sites shown by our evolutionary comparison to be under control of CRP). An evolutionary analysis of the population of sites that bind strongly according to `crp-core-only` and also lie in coding regions is instructive: no significant position-dependent mutation pattern is seen, further confirming that these sites are spurious (see Supporting Information).

5.4 Discussion

Novel CRP sites are probably functional. Attempts to describe the sequence-specific DNA-binding affinity of CRP by a linear additive energy model always lead to the prediction of many more strong binding sites (and regulated genes and operons) than are verified by direct experimental methods. We used the DPInteract site compendium (11) to demonstrate this, but the same conclusion would have been reached had we used other databases such as RegulonDB (12) (see Supporting Information). It is important to obtain independent information about the status of the many predicted novel binding sites: if they are spurious, the validity of the linear additive binding energy model is called into question; if they are real, the validity of that model is reaffirmed in a challenging context and something new is learned about how CRP functions.

This issue could in principle be addressed experimentally, although that approach has drawbacks: *in vivo* tests may fail to expose the role of a given CRP motif because the test conditions are wrong, while *in vitro* tests such as gel shifts or DNAase footprinting can score pseudosites as functional CRP binding sites. We have instead sought evidence of their functionality via a computational study of binding site evolution between closely-related bacterial species. We have presented several lines of evidence suggesting that the novel sites are mostly functional, with a likelihood that increases with the strength of the predicted binding (as measured by the CRP energy matrix). In increasing order of importance, they are:

1. The novel CRP sites are overwhelmingly located in non-coding regions, exactly where regulatory sites should lie. Since the *E. coli* genome is only 15% non-coding and since spurious sites should be randomly located, this is an improbable chance occurrence. This observation is independent of evolutionary considerations.

2. If a novel strong *E. coli* site lies in an intergenic region with an *S. typhimurium* ortholog, an orthologous site pair can be defined by aligning the two regions. The difference in predicted binding energy between the two sites is systematically less than would be expected on the hypothesis of random mutations.

3. Within a population of orthologous pairs of novel strong binding sites, the probability of mutation depends strongly on position within the site; random mutations would have led to a position-independent profile. The actual position-dependent pattern is consistent with the way CRP is known to contact the DNA.

The cross-species comparison shows that populations of orthologous sites defined by strong predicted CRP binding energy have a non-random mutation pattern consistent with conservation of CRP binding energy. Conservation makes sense if the sites are functional for CRP and this evidence suggests that the primary determinant of functionality is the CRP binding energy of the site sequence. We have focused on sites with orthologs, but that was a device to select a sub-population of sites on which functionality could be observed via its effect on mutations. Therefore, although the evolutionary evidence applies directly only to sites with orthologs, we suggest that strong sites in intergenic regions without orthologs are also likely to be functional for CRP.

The specific outcome of these considerations is a list of computational predictions of novel CRP binding sites in *E. coli*. We include all sites with $E < 9$ (the cutoff at which the mutation profile of the site pair population becomes indistinguishable from that of the starting databases; see Supporting Information) whether or not they have a companion species ortholog. A short list, generated with the stringent cutoff $E < 4$, is given in Table 6. The long list, containing more than 190 novel sites, is available on-line (see Software). The genes downstream from these sites would be interesting targets for investigation of the influence of CRP

Table 6: List of genes with putative CRP sites upstream of the operon; an energy cutoff of 4.0 was used. Where an orthologous site exists in *S. typhimurium*, the energy of that site is also given. An extended list is available on our Web site; see Software.

Upstream site	**Downstream gene(s)**	**Aligned site**
2.59	tsr (b4355); yjiY (b4354)	-
2.63	b1904 (b1904)	-
3.14	yjcB (b4060); yjcC (b4061)	3.33
3.19	nupG (b2964)	9.14
3.42	mtlA (b3599); yibI (b3598)	-
3.43	b1458 (b1458)	-
3.49	tnaL (b3707)	-
3.66	qseA (b3243); yhcR (b3242)	-
3.74	yeaA (b1778); b1777 (b1777); gapA (b1779)	5.87
3.77	ydeA (b1528)	-
3.77	hpt (b0125); gcd (b0124)	4.47
3.83	ycfQ (b1111); ycfR (b1112)	6.70
3.87	proP (b4111)	-
3.94	ygjG (b3073); aer (b3072)	-

on their expression levels. The new genes do not, by and large, appear in the most comprehensive databases of regulatory information (see Supporting Information), suggesting that their regulation by CRP is subtle. On the other hand, the evolutionary evidence that such effects are important to the organism is strong. Understanding how and why will be an enlightening enterprise.

Binding energy, not sequence, is conserved. Figure 1 shows that the binding energies of sites in *E. coli* correlate well with the binding energy of independently aligned sites in *S. typhimurium.* Figure 2 shows that this is not, however, due to strict sequence conservation: outside the eight core positions, the individual site pairs differ significantly at the sequence level. Others (20) have also noted the position-dependent mutation rate between orthologous binding sites. We have observed that this position dependence becomes more and more pronounced as the binding energy computed from the `CRP-naive` matrix becomes stronger. This suggests that the underlying cause of the mutation pattern is conservation of site binding energy between the two species. It also suggests that the matrix model binding energy is closely related to the actual *in vivo* binding energy. This agrees with the prediction made by Berg and von Hippel on theoretical grounds (16).

Lessons for practical motif hunting. While these observations are useful for considering the actual nature of genomic binding of transcription factors, we can also try to draw some conclusions regarding the utility of various types of computational searches for novel transcription factor binding sites on a whole-genome scale.

A promising technique for locating functional regulatory regions *de novo* identifies as putative binding sites elements of genomic sequence conserved between two or more closely related species (6, 21). However, we have found that sites may diverge at the sequence level without significantly changing the binding energy, as shown in Figures 1 and 2. Indeed, our comparisons of *E. coli* and aligned *S. typhimurium* regions containing predicted CRP binding sites suggest that many of the sites would not be identifiable on the basis of local sequence conservation.

Another prevalent technique for binding site searches is to use a "consensus sequence", consisting of the bases that appear most frequently in the list of known sites. In the case of CRP, searches for even the very general consensus sequence `GTGANNNNNNTCAC` would fail to discriminate between our novel (putatively functional) sites and nonfunctional sites, due to the lack of weighting from the flanking sequence. Searches with more restricted sequences would necessarily recover only a subset of our novel sites.

These observations suggest that a full matrix model will be needed to give optimum discrimination in separating actual binding sites from the rest of the genome. Unfortunately, there is usually only a limited number of known sequences from which to infer the matrix and a large error in its construction. Our results suggest that statistics could be improved by using aligned sites from orthologous intergenic regions as additional inputs. Beyond that, it may be fruitful to pursue optimization methods for tuning matrices to better represent TF binding. The results of using one such method (developed by us in unpublished work (19)) to improve the LacI energy matrix were presented earlier in the paper. An independent study of CRP binding in *E. coli* that uses an optimization procedure (QPMEME) to refine the energy matrix has recently appeared (15). It also finds many novel predicted CRP sites and addresses the issue of whether they are real on the basis of their positions relative to transcription initiation sites. The quite interesting results of applying our species comparison assessment of functionality to the QPMEME site list are presented in Supporting Information. A comparative study of the performance of different optimization methods will be presented elsewhere.

Concluding Remarks. We have demonstrated that many computationally identified CRP binding sites have non-random mutation patterns, strongly suggesting that they are functional. Sites that have strong predicted binding energy, but no ortholog in *S. typhimurium* are also likely to be actual CRP sites, although our method gives no direct evidence of their functionality. Combined, we find that there are more than 190 novel CRP binding sites in the *E. coli* genome, none of them known from experimental evidence. While this is important information about CRP itself, we would like to emphasize the equally important conclusion that the linear energy matrix model for TF binding works well in a case where it had been thought to generate far too many false positives. It will be interesting to explore this issue across a wider range of transcription factors and organisms.

We thank Erich M. Schwarz, Paola Oliveri, and Saeed Tavazoie for useful discussions. We also thank Nikolaus Rajewsky and Leonid Kruglyak for careful reading of the manuscript and helpful suggestions. Finally, we thank Eric H. Davidson, R. Andrew Cameron, and the Beckman Institute Center for Computational Regulatory Genomics at Caltech for access to their computational resources, supported by NIH grant RR15044. CTB was supported by NIH grant GM61005 to E. H. Davidson.

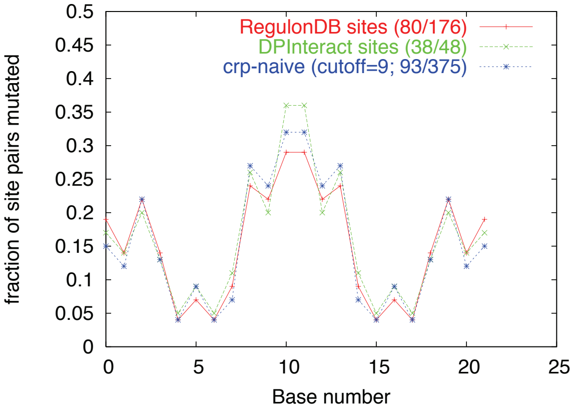


Figure 5.3: Cross-comparison of databases.

5.5 Supporting Information

Using RegulonDB instead of DPInteract. There are two extant collections of CRP binding sites: RegulonDB and DPInteract. We used the DPInteract collection in this paper because all of the sites were tested *in vitro*, while some of the sites in RegulonDB were taken from computational studies. Here we show that our results are independent of the starting database.

The DPInteract list of sites contains 48 distinct entries and the RegulonDB list of sites contains 151 distinct entries. A total of 41 sites lie in the intersection of these two lists. Thus, RegulonDB contains most of the sites in DPInteract, plus many additional sites. To examine the use of RegulonDB instead of DPInteract, we used `openfill` (see Methods) to construct a matrix from the RegulonDB site list (available on the Web site in correct format) and calculated the statistics of site energies in the genome with `scangen`. The results are displayed in Table 7. They demonstrate that using the RegulonDB list as the initial site list does not result in a matrix that preferentially picks out the initial sites. Thus, there is no gain in specificity from using RegulonDB sites for the initial site list.

We then compared the pattern of conservation in sites from the DPInteract list with the pattern of conservation of sites from the RegulonDB database, and cross-compared with sites isolated using the `CRP-naive` matrix for $E < 9$. Figure 3 demonstrates that the three site lists have nearly identical patterns and levels of conservation, which is particularly striking given that there are more than twice as many sites in the `CRP-naive` generated list as in the RegulonDB list. This suggests that both the DPInteract and RegulonDB lists are, at best, subsets of the list of all CRP binding sites in the genome.

Comparison with the QPMEME approach. Recently, Djordjevic, Sengupta and Shraiman [15] presented a novel algorithm for creating transcription factor binding matrices called QPMEME (Quadratic Programming Method of Energy Matrix Estimation). This algorithm was used to generate a list of likely

Cutoff	# sites	coding	intergenic	RegulonDB sites
5.00	28	4%	96%	11/176
7.00	109	8%	92%	33/176
9.00	379	24%	76%	83/176
11.00	1680	53%	47%	119/176
15.00	37269	72%	28%	161/176

Table 7: Scangen list for CRP binding, using a matrix built from the RegulonDB site list.

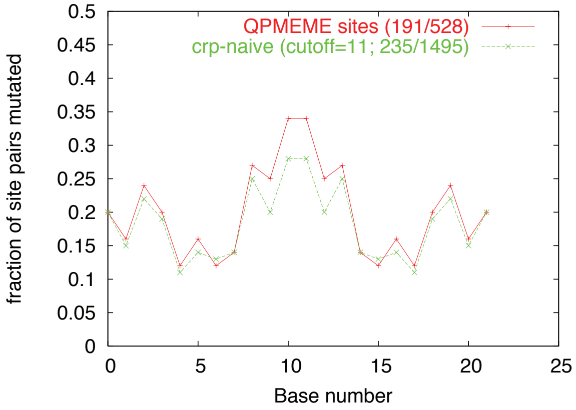


Figure 5.4: Cross-comparison of computationally generated site lists.

CRP binding sites in the *E. coli* genome. It is instructive to compare their results with ours.

In the QPMEME-generated list of sites for CRP, there are 528 distinct entries. We calculated the per-position mutational bias for the 191 *E. coli* sites in their list for which there are orthologs in *S. typhimurium.* We compared this with the per-position mutational bias in the list of *E. coli* sites generated by `CRP-naive` with a cutoff of $E < 11$. The latter contains 1495 sites, of which 235 have clear orthologs in *S. typhimurium.* The results, shown in Figure 4, are nearly identical, suggesting that both lists contain equivalent site populations. However, the lists are not numerically equivalent: only 413 of the 528 sites in the QPMEME list are also in our $E < 11$ list, leaving more than 1075 sites in only the $E < 11$ list and 115 sites in only the QPMEME list.

These observations suggest that both lists contain significant numbers of good CRP sites without being inclusive of all such sites. In fact, both the QPMEME approach and our approach seem to have a significant false negative rate: there are 58 sites in RegulonDB that do not appear in our list even at an $E < 15$ cutoff, although at that cutoff our list contains all but two of the QPMEME sites. This, in turn, suggests that we do not yet have a complete picture of CRP binding in the *E. coli* genome.

Comparison of `CRP-core-only` sites. In Table 5, we show that the matrix `CRP-core-only` picks out many sites in coding regions. In the text, we interpreted this as evidence that these sites are nonfunctional

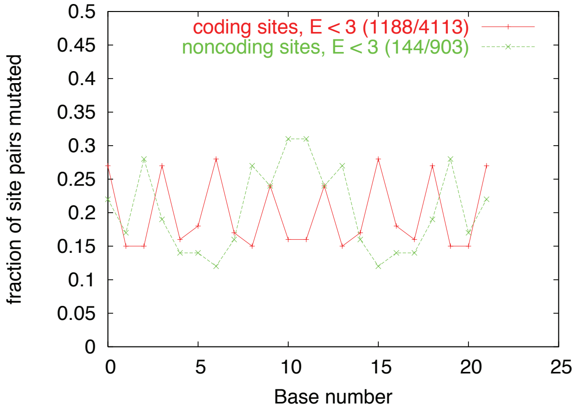


Figure 5.5: Positional mutation bias in sites found with the `CRP-core-only` matrix; comparison between coding region sites and noncoding region sites.

and that the bases flanking the eight core bases lend important discriminatory power to the `CRP-naive` matrix. However, these sites could in principle be valid CRP binding sites and it is interesting to test this using our species-comparision methodology. To do this, we aligned all *E. coli* coding regions with the orthologous *S. typhimurium* coding regions using CLUSTALW (see Methods), and compared the sites in *E. coli* with aligned sites in *S. typhimurium*. The results are displayed in Figure 5, using the same format as in Figure 2. The population of noncoding region sites displays the familiar position-dependent mutational bias (quantitatively very similar to the $E < 11$ curve of Figure 2). On the other hand, the population of coding region sites displays a completely different pattern: periodic, with a period of three bases, but otherwise position-independent. This periodic bias probably arises from conservation of amino acid sequence between orthologous genes but, in any event, shows no hint of the characteristic pattern of mutation due to CRP binding. This strengthens our argument that the sites identified by `CRP-core-only` in coding regions of *E. coli* are unlikely to be real CRP binding sites.

Bibliography

Ptashne, M. (1992) *A Genetic Switch*, Blackwell Scientific Publications.

Sinha, S. & Tompa, M. (2003) *Nucleic Acids Res.* **31**(13), 3586-3588.

Rajewsky, N., Vergassola, M., Gaul, U. & Siggia, E.D. (2002) *BMC Bioinformatics* **3**(1), 30.

Lenhard, B., Sandelin, A., Mendoza, L., Engstrom, P., Jareborg, N. & Wasserman, W.W. (2003) *J. Biol.* **2**(2), 13.

Markstein, M., Markstein, P., Markstein, V. & Levine, M.S. (2002) *Proc. Natl. Acad. Sci.* USA **99**(2), 763-768.

McCue L.A., Thompson W., Carmack C.S. & Lawrence C.E. (2002) *Genome Res.* **12**(10), 1523-32.

Berg, O.G. & von Hippel, P.H. (1988) *Trends Biochem. Sci.* **13**(6), 207-211.

Benos, P.V., Lapedes, A.S. & Stormo, G.D. (2002) *Bioessays* **24**(5), 466-475.

Benos P.V., Bulyk M.L. & Stormo G.D. (2002) *Nucleic Acids Res.* **30**(20), 4442-51

Cashel, M., Gentry, V.J., Hernandez, V.J. & Vinella, D. (1996) in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, ed. Neidhardt, F. C. (Am. Soc. Microbiol., Washington, DC), pp. 1458 - 1496.

Robison, K., McGuire, A.M. & Church, G.M. (1988) *J. Mol. Biol.* **284**(2), 241-254.

Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millan-Zarate. D., Diaz-Peredo, E., Sanchez-Solano, F., Perez-Rueda. E., Bonavides-Martinez, C. & Collado-Vides, J. (2001) *Nucleic Acids Res.* **29**(1), 72-74.

Berg, O.G. & von Hippel, P.H. (1987) *J. Mol. Biol.* **193**(4), 723-750.

Tan, K., Moreno-Hagelsieb, G., Collado-Vides, J. & Stormo, G.D. (2001) *Genome Res.* **11**(4), 566-584.

Djordjevic, M., Sengupta, A.M. & Shraiman, B.I. (2003) *Genome Res.* **13**(11), 2381-2390.

Berg, O.G. & von Hippel, P.H. (1988) *J. Mol. Biol.* **200**(4), 709-723.

Rajewsky, N., Socci, N.D., Zapotocky, M. & Siggia E.D. (2002) *Genome Res.* **12**(2), 298-308.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. & Thompson, J.D. (2003) *Nucleic Acids Res.* **31**(13), 3497-3500.

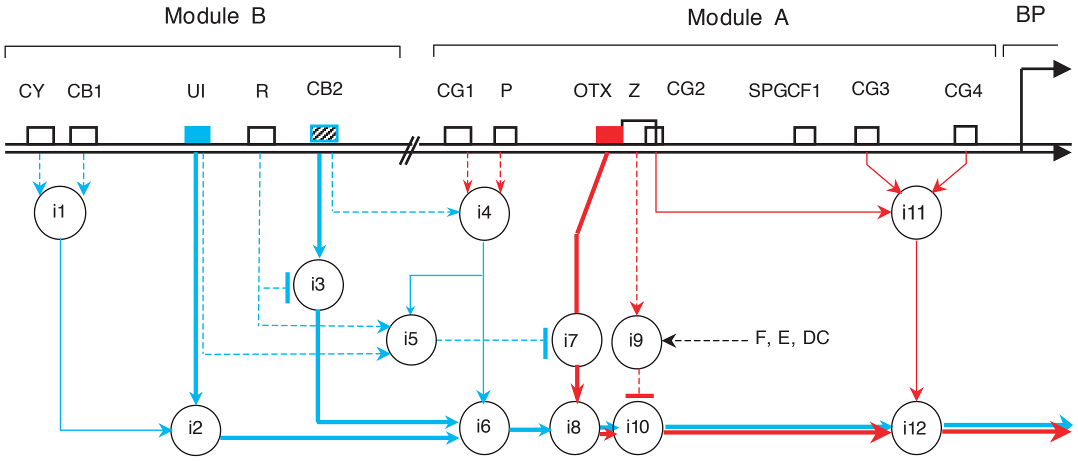
Brown, C.T. & Callan, C.G., Jr. *unpublished.*

Moses, A.M., Chiang, D.Y., Kellis, M., Lander, E.S. & Eisen, M.B. (2003) *BMC Evol Biol.* **3**(1), 19.

van Nimwegen, E., Zavolan, M., Rajewsky, N. & Siggia, E.D. (2002) *Proc. Natl. Acad. Sci.* USA **99**(11), 7323-7328.

Chapter 6

Conclusion



if CY & CB1	i1 = 1	if i5 = 0	i7 = OTX(t)
else	i1 = 0.5	else	i7 = 0
			i8 = i6 + i7
	i2 = i1 • UI (t)		
		if (F or E or DC) & Z	i9 = 1
		else	i9 = 0
if R	i3 = CB2(t)		
else	i3 = k • CB2(t) (1<k<2)	if i9=1	i10 = 0
		else	i10 = i8
if P & CG1 & CB2	i4 = 2		
else	i4 = 0	if (CG 2 & CG3 & CG4)	i11 = 2
		else	i11 = 1
if UI (t)>threshold & R & i4‡0	i5 = 1		i12 = i11• i10
else	i5 = 0		
	i6 = i4 • (i2+i3)		

Figure 6.1: Computational logic model for the BA region of the *endo16 cis*-regulatory system, reproduced from (Yuh et al., 2001). As in the logic diagram for *cyIIIa* (Figure 4.4.6), red lines indicate module A function and blue lines indicate module B function. Heavy solid lines trace driver function, while light solid lines indicate amplifier function. Dashed lines indicate Boolean logic interactions. (Reproduced with permission of the Company of Biologists.)

The *endo16* and *cyIIIa cis*-regulatory regions are extensively investigated regions that regulate expression of two differentiation genes in the developing sea urchin embryo. The *endo16* region contains over 30 binding sites for 15 different species of factors, and *cyIIIa* contains 15 binding sites for 9 distinct factors. Each individual binding site in the A and B modules of both of these regions has been mutated to abolish the protein-DNA interaction at that site, and from that we know that each site is causally involved in the regulatory function of the region. Moreover, the function of most binding sites in these *cis*-regulatory regions has been comprehensively analyzed and the known binding sites are, in aggregate, sufficient to explain the regulatory function entirely. We now have an excellent opportunity to compare and contrast the regulatory functions encoded by these two well-studied regions.

Function type	*endo16* site/interaction	*cyIIIa* site/interaction
Drivers	Otx, CB2, UI sites	P1, Runt sites
Silencers	F, E, DC modules (i10)	-
Combinatorial (short-range repressors)	R site (i3)	P3A2 site (i1) Myb, Z12 sites (i5)
Combinatorial (AND operators)	CY, CB1 sites (i1) CB2, CG1, P sites (i4) UI, R sites (i5) CG2, CG3, CG4 sites (i11)	Oct-1, CBF, TEF-1 sites (i4)
Amplifiers	CY, CB1 (i2) CG2, CG3, CG4 (i12)	Oct-1, CBF, TEF-1 sites (i3) distal SpGCF1 sites (i8)
Communicators	Z (i9)	Oct-1, CBF, TEF-1 sites (i4)
Module linkers	CB2, CG1, and P sites (i6)	-
Switch	UI, R, P, CG1, CB2 (i7)	-
Module output summation	i8	i2, i7

Figure 6.2: Classification of *cis*-regulatory functions present in the *endo16* and *cyIIIa* *cis*-regulatory regions. The nomenclature of (Istrail and Davidson, 2005) is here extended to cover all of the observed interactions present in *endo16* and *cyIIIa*; specifically, the summation functions (i8 in *endo16* and i2 and i7 in *cyIIIa* and the switch function (i7 in *endo16*) were added. The common DNA-looping function mediated by SpGCF1 binding sites is omitted except in the case of the distal cluster in *cyIIIa*, which exerts a constant amplifying function.

In both regions, the effect of one binding site mutation is not necessarily separable from the effect of another binding site mutation: these sites evidently work in combination to direct transcriptional output. However, despite the number and complexity of the protein-DNA interactions in these two *cis*-regulatory regions, the functional consequences of binding site presence or absence can be described with relatively simple logic diagrams. These logic diagrams encapsulate the essential interaction logic between binding sites in these *cis*-regulatory regions, despite containing no explicit information about protein-protein interactions.

The success of this approach demonstrates that the specificity of gene regulation in these two regulatory regions rests primarily in the DNA-protein interactions. To the extent that this generalizes to many developmental regulatory regions, the identity of transcription factors that bind in a *cis*-regulatory region together with the arrangement of binding sites in that region determines the function of the region (reviewed in (Davidson, 2006)). Off-DNA interactions such as phosphorylation and co-factor proteolysis generally inhibit DNA binding altogether, as with P3A2. (One significant exception is seen in signalling switches such as Su(H), in which the DNA-binding factor changes function from a repressor to an activator when a co-factor is present.) This lets us at least attempt to understand *cis*-regulatory regions solely as a function of the binding sites present in the region.

This approach is convenient, because it is increasingly clear that the *endo16* and *cyIIIa* *cis*-regulatory regions are not exceptional in their complexity. Many genes possess regulatory regions that are significantly more complicated than *cyIIIa* and *endo16*, with many distinct modules, each driving function independently, spread across 10s of kilobases of DNA. Our ability to understand the regulatory interactions mediated by the relatively simple *endo16* and *cyIIIa* regulatory regions is critical to understanding how these even more complex regions function.

We are still unable to "read" *cis*-regulatory sequence by inspection, i.e. without experimental analysis. Despite a plethora of information about individual binding sites and the identification of many regulatory regions, we cannot reliably locate binding sites within regulatory regions, much less decode the logical functions performed by these binding sites. It is clear that different arrangements of binding sites can result in similar output. For example, the *endo16* UI site alone reproduces the B module time course and spatial effect, albeit at a lower absolute level of expression; and the P3B, P4, and P5 sites in the *cyIIIa* A module function even when placed in a completely artificial background sequence ((Yuh et al., 2001); Chapter 4 of this thesis). This many-to-one mapping of function, together with the significant degeneracy of transcription factor binding, effectively prohibits us from generalizing from knowledge of existing regulatory regions.

Even though we cannot predict *cis*-regulatory function from sequence alone, we can identify a minimal set of logic functions encoded by regulatory regions. This will help us to identify likely interactions and may eventually let us construct a lexicon of transcription factor function. (Istrail and Davidson, 2005) initiate such an effort based primarily on data from *endo16*; below, we discuss their classification in light of the new analysis of *cyIIIa*.

Figure 6.2 contains a breakdown of the functions encoded by the *endo16* and *cyIIIa* regulatory regions.

The majority of logic interactions in both *endo16* and *cyIIIa* are combinatorial: most binding sites function either as short-range repressors or as AND operators. This is not unexpected, given the role that regulatory regions play in integrating diverse inputs.

The second most prevalent type of logic interaction is the driver interaction. In *cyIIIa* and *endo16*, there are five distinct driver interactions: P1 and Runt in *cyIIIa*, and Otx, UI, and CB2 in *endo16*. These driver interactions specify at what time and in which territory the regulated gene should be transcribed.

Repressors are responsible for confining expression to the appropriate territories. In *endo16*, repression acts through the Z "communicator" site in the A module (discussed below), and no spatial repressors bind in the BA modules directly. In *cyIIIa*, the P3A2 repressor binds directly to the A module, while the Myb and Z12 repressors bind in the B module. Despite this difference in organization, repression acts as a Boolean function in both *endo16* and *cyIIIa*: when the repressors are present and bound, the *cis*-regulatory region does not drive expression. Because the *endo16* repressor modules (the DC, E, and F modules) are external to the B module which contains the driver functions being repressed, the repression is not short-range and the function is classified as a silencer. In contrast, the *cyIIIa* repressor sites act only on local drivers, as does the *endo16* R repressor, so they are clearly short-range repressors.

Binding sites for transcription factors that mediate amplifier functions also act throughout both of these regulatory regions. Some amplification works within modules, such as the combined effect of the CY and CB1 sites in *endo16*. Other amplification works as a general amplifier of function, such as the CG2/CG3/CG4 effect on the combined output of the A and B modules in *endo16*. These amplifier functions are not time dependent, but rather multiply the output of driver sites by a constant factor. The only exception to this is in the module A amplification mediated by the Oct-1, CBF, and TEF-1 sites, which serve both as amplifiers (at 30 hours) and communicators (at 30 and 50 hours; see below).

SpGCF1 binding sites are prevalent throughout both *cis*-regulatory systems. Removal of SpGCF1 sites results in a decrease of transcriptional activity in both *endo16* and *cyIIIa*; because SpGCF1 multimerizes *in vitro*, we believe that SpGCF1 also multimerizes *in vivo* and is responsible for enhancing communication between *cis*-regulatory modules, thereby increasing overall transcriptional output.

There are two rarer types of interactions present in the *endo16* and *cyIIIa* regulatory regions.

One type of interaction that is shared by *cyIIIa* and *endo16* is the "communicator" function (Istrail and Davidson, 2005). Here, one or more binding sites mediates an interaction external to their module; elimination of these sites negates the effect of the external interaction. In *cyIIIa*, the Oct-1, CBF, and TEF-1 binding sites are individually required for module B output; constructs lacking any of these sites work equivalently to a construct containing only the A module alone. In the *endo16* regulatory region, the Z site mediates repressive inputs from the DC, E, and F modules: elimination of this site is equivalent to removing these three modules entirely. Thus both the Oct-1/CBF/TEF-1 and the Z interactions mediate communicator functions. In both cases, the communicator function has no apparent spatial specificity: expansion of the *cyIIIa* B module function through deletion of repressor sites expands the expression of reporter constructs to the entire embryo, while the *endo16* DC, E, and F modules repress reporter expression throughout much of the embryo. Therefore both communicator functions seem to serve as architectural aids rather than as repositories of additional spatiotemporal regulatory information. These sites may generally serve to target the distal CRMs to a specific basal promoter, a conjecture supported by the proximity of the sites to the basal promoters in both *endo16* and *cyIIIa*.

Another type of interaction, the linker interaction, is present only in the *endo16* regulatory region; no such interaction occurs in *cyIIIa*. The *endo16* linker interaction is mediated by the CB2 site in the B module of *endo16*, and the P and CG1 sites in the A module. Deletion of any one of these three sites eliminates the entire effect of the B module on transcriptional output. While superficially similar to the communicator function described above, these two interactions work very differently: the communicator function is *required* for the *cyIIIa* B module function, no matter where the B module is physically located. In contrast, the B module sites function normally when moved closer to the basal promoter, suggesting that the CB2-P-CG1 function is distance-related – that is, it is a linker function. One explanation for these logical functions is that the communicator function is probably a matter of biochemical necessity by which one protein communicates with the BTA via specific interactions with others. The linkage site simply serves an architectural function by linking the *endo16* B module physically to the basal promoter.

The communicator and linker interactions most likely serve as a form of traffic management, in which the effect of more distal regulatory elements is integrated at the basal promoter. Such interactions are more difficult to identify than driver or repression interactions, which can be detected by the expansion or contraction of reporter expression from mutated constructs. The removal of linker or communicator interactions may mimic the effect of driver or repressor removal, e.g. the removal of the Z binding site in *endo16* is equivalent to the removal of early repression. This could lead to confusion about which sites and regulators confer proper spatial and temporal expression within a regulatory region.

The change from the early effect of the Oct-1, CBF, and TEF-1 factors on module A function in *cyIIIa* is the only effect identified in either regulatory region that may involve a modification to the function of a transcription factor during embryogenesis. Between 30 and 50 hours post-fertilization, the Oct-1, CBF, and TEF-1 factors cease to amplify module A output; this effect must be dependent on some feature of the P1 driver, because it is the only driver affected by these factors. This is unlike all of the other functions revealed by binding site deletion, which can be understood as either a Boolean effect based on the presence or absence of the bound factor (all repressive interactions as well as the amplifier, linker, and communicator functions) or a concentration-dependent effect (the driver interactions). Without further investigating the time course of the A module, or identifying the gene encoding the P1 gene, we cannot say whether or not the switch in function depends on a modification to P1 or a change in the amount of P1 bound to the site.

cis-regulatory analysis is critical for understanding the role that each individual transcriptional regulatory factor plays in regulating the transcription of downstream genes. Driver and repressor functions are relatively easy to identify, once the identity of regulators are known. However, without a detailed understanding of how linker and communicator functions work in a given regulatory region, we cannot reliably determine the effect of the drivers and repressors on that region.

Another important reason to subject *cis*-regulatory regions to the kind of detailed analysis discussed above is that we ultimately would like to be able to "read" regulatory function by inspecting regulatory sequence. Experimental analysis is always going to be slow and difficult; computational aids could significantly speed our ability to build gene regulatory networks. Yet without more examples we cannot hope to build a general understanding of the regulatory code that will enable us to read regulatory sequence computationally.

Bibliography

Davidson, E., 2006. *The Regulatory Genome*. Academic Press.

Istrail, S. and Davidson, E., 2005. Logic functions of the genomic cis-regulatory code. *Proc Natl Acad Sci U S A*, **102**(14):4954–9.

Yuh, C., Bolouri, H., and Davidson, E., 2001. Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. *Development*, **128**(5):617–29.

Appendix A

Statistics and Scoring for `paircomp` comparisons

A.1 Introduction

This is a miscellaneous collection of the math "theory" behind paircomp/seqcomp-style fixed-width-window matching. The first section, on Monte Carlo sequence matching probabilities, was mostly written back in 2000, at the request of my advisor, Eric Davidson; the second section, on building a score for matches, was written in late 2003, at the request of Erich Schwarz.

Tristan De Buysscher is the author of the seqcomp program that I used during this period, and my co-conspirator in this matter.

A.2 Monte Carlo sequence matching probabilities

The problem of finding a match of exactly M base pairs of a particular W-mer reference sequence in some subsequence of L equiprobably chosen bases is equivalent to the following problem: pick W balls from a large collection of red and blue balls, with probability $p = \frac{1}{4}$ of picking a blue ball and probability $q = \frac{3}{4}$ of picking a red ball; do this picking once if $L = W$, or repeat it if $L > W$ until one has done independent pickings a total of $(L - W + 1)$ times; and, in each individual picked set, look for exactly M blue balls of the W total chosen.

The probability of finding exactly M blue balls out of the W chosen is a simple binomial calculation:

$$P(M, W) = \binom{W}{M} p^M q^{W-M}$$

The corresponding distribution peaks at the mean, Wp, and has variance Wpq.

When repeating L times ($L >> W ==> L - W + 1 \approx L$) we can ask for the probability of getting one or more candidate matches by calculating the Poisson probability of exactly 0 matches and then subtracting it from 1:

$$\mathcal{P}_L(E) = 1 - \exp(-\lambda(L))$$

where E is the occurrence of one or more matches, and the formula is for the Poisson distribution at $x = 0$. The Poisson parameter λ is a function of the size of the sequence L (equivalently, the number of candidate matches generated), and can be chosen by setting λ equal to the mean number of events. For a binomial distribution bin(n, p) the mean is np, and if we require that this match the mean of the Poisson distribution, we find that

$$\lambda(L) = Lp$$

and then

$$\mathcal{P}_L(E) = 1 - \exp(-Lp)$$

By way of intuitive confirmation, note that the distribution (figure) asymptotically approaches 1, so

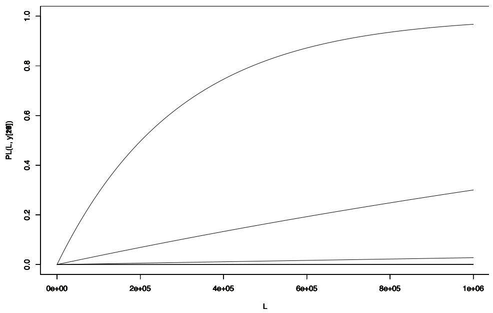


Figure A.1: Probability of seeing one or more motif matches for a given sequence length across matches of 15/20 to 20/20.

that the probability of generating at least one match converges to one as the number of tries approaches infinity; as well, the probability of getting a match in few tries (given a rare event, equivalently a small p) is approximately 0.

A.2.1 Distribution of maximum matches

Now suppose we have an ensemble of sequences of length L and we would like to describe the distribution of *maximum* matches to a W-mer on members of the ensemble. That is, given a particular W-mer matched against every W-length subsequence of each sequence in the ensemble, what is the distribution of best matches to the W-mer across the ensemble?

To describe this in terms of the probability distribution derived above, we need only consider two types of events: the event E in which at least one subsequence has M matches, and the event NE in which no subsequences have (M + 1) or more matches. Then the probability that the best match to a subsequence of an L-length sequence is M is

$$W_L(E)(1 - W_L(NE))$$

Now, $W_L(E) = \mathcal{P}_L(E)$, and $W_L(NE) = \mathcal{P}_L(NE)$ which can be generated by replacing the probability p of generating M matches with the probability $p_W M = \sum_{i=M+1}^{W} bin(W, i)$, that is, the probability that no more than M matches are generated in a given attempt.

The resulting probability distribution is graphed for a fixed $L = 1e6$, below.

A.2.2 Distribution of best matches in diverging sequences

Suppose that you have a sequence of L bases, each of which is mutated with a probability p (which can be related to a Poisson rate of mutation and a waiting time via an exponential, as above). The problem now changes so that you start with a full set of matches and then randomly alter the selection of matches with

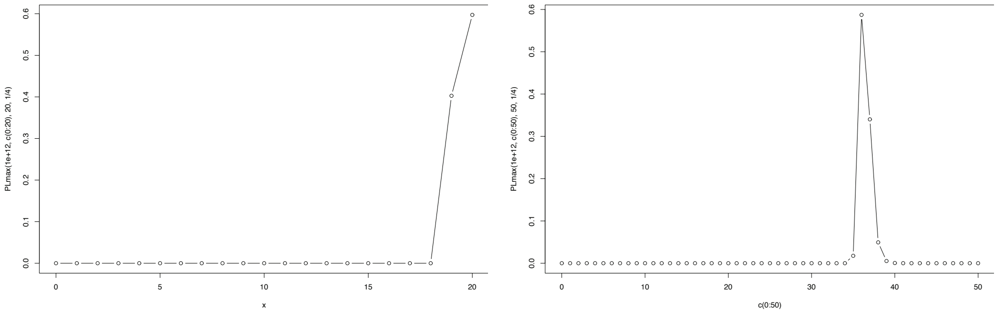


Figure A.2: Distribution of best matches for windows of size 20 and 50, for sequences of length 10^6.

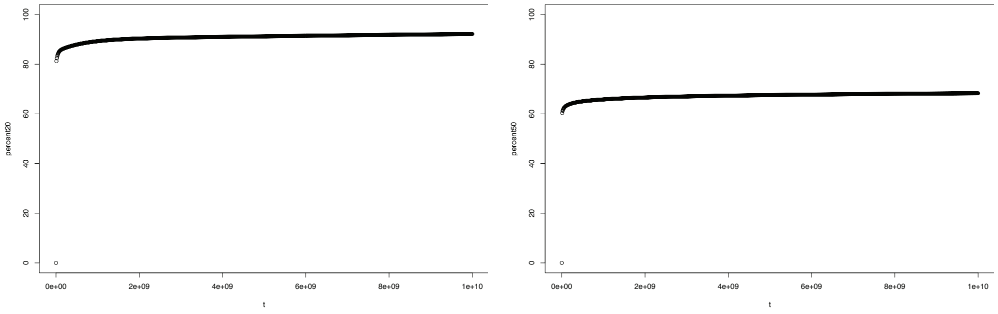


Figure A.3: Expected best matches, according to sequence length, for two different match-window sizes: top = 20, bottom = 50.

some probability. Accounting for back mutations, then, the probability that any given base still matches is

$$P(p) = (1-p) + \frac{p}{4} = 1 - \frac{3p}{4}$$

– the probability of no mutation plus the probability of back mutation. Checking boundary conditions reveals the expected results: for $p = 0$ (no mutation), $P = 1$; for $p = 1$ (no unchanged bases), $P = \frac{1}{4}$, random (as discussed above).

A.3 Scoring matches

Assume a match between two sequences a, b of length N. The probability of these two sequences being generated randomly (hypothesis R) is

$$p(a, b|R) = \prod_i q_{a_i} \prod_j q_{b_j}$$

where q_{a_i} is the probability of the i'th nucleotide in sequence a; for a equiprobable distribution of bases,

$$q_{a_i} = q_{b_j} = p_A = p_C = p_G = p_T = \frac{1}{4}$$

The probability of these two sequences being aligned (hypothesis M) assuming independence between bases is that

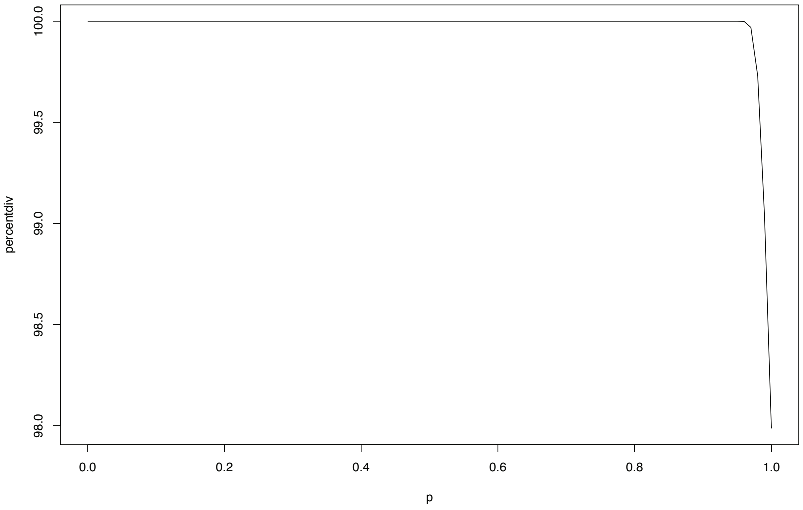
$$p(a, b|M) = \prod_i p_{a_i, b_i}$$

which must be chosen according to some model of nucleotide matching. For comparative sequence analysis of noncoding regions, this probability corresponds to our belief that an identity between two nucleotides indicates conservation for functional reasons as opposed to accidental failure to diverge. Since we have no reason to bias ourselves for or against any particular nucleotide identity, we can pick a single parameter λ to represent the likelihood for all four bases that the nucleotide match is due to conservation, which gives us

$$\begin{aligned} p_{a_i, b_i} &= \delta_{a_i, b_i} \left(\frac{1}{16} + \frac{3\lambda}{16} \right) + (1 - \delta_{a_i, b_i}) \left(\frac{1}{16} - \frac{\lambda}{16} \right) \\ &= \delta_{a_i, b_i} \left(\frac{1 + 3\lambda}{16} \right) + (1 - \delta_{a_i, b_i}) \left(\frac{1 - \lambda}{16} \right) \end{aligned}$$

where δ_{a_i, b_i} is the identity matrix, 1 when $a_i = b_i$ and 0 otherwise, and $\lambda \in [0, 1)$. This can be broken down into two probabilities, p_{match} and p_{mismatch}:

$$p_{\text{match}} = \frac{1 + 3\lambda}{16}, p_{\text{mismatch}} = \frac{1 - \lambda}{16}$$



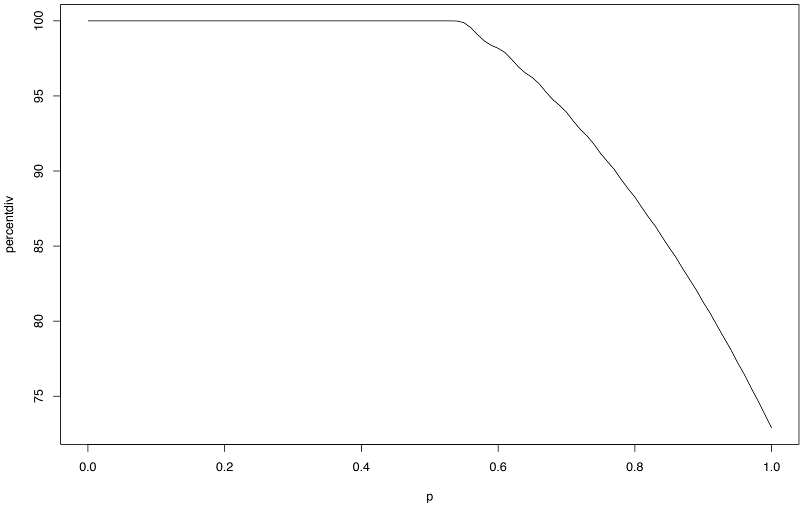


Figure A.4: Expected best matches for a sequence of length 10^{12}, with windowsize 20 (top) and 50 (bottom), for different divergences. Divergences are measured here as values of p, the probability that any given base pair is mutated.

Note that for $\lambda = 0$ – no identity due to conservation – these joint probabilities equal the independent probabilities, and for $\lambda = 1$ the probability of any given non-identical match is 0, which will certainly be counter to observation for any inhomogeneous sequence regions.

Given these two probability measures, we can now construct a log-ratio score according to (Durbin et al., 1999), p13-14:

$$\begin{aligned} s(a,b) &= \log \frac{p(a,b|M)}{p(a,b|R)} \\ &= \log \left(\prod_i \frac{p_{a_i,b_i}}{q_{a_i} q_{b_i}} \right) \\ &= \sum_i \log \frac{p_{a_i,b_i}}{q_{a_i} q_{b_i}} \\ &= \sum_i \log p_{a_i,b_i} - \log q_{a_i} q_{b_i} \end{aligned}$$

This can easily be calculated for any particular value of λ; for M nucleotide matches of an N-length sequence,

$$s(a,b) = M \log p_{\text{match}} + (N - M) \log p_{\text{mismatch}} - \sum_i \log q_{a_i} q_{b_i}$$

In the simplest case, for a uniform distribution of nucleotides in the background sequence, $q_{a_i} = q_{b_i} = \frac{1}{4}$,

$$\sum_i \log q_{a_i} q_{b_i} = \sum_i 2 \log \frac{1}{4} = -4N$$

and so

$$s(a,b) = M \log p_{\text{match}} + (N - M) \log p_{\text{mismatch}} + 4N$$

This scoring technique has been implemented in Python and is available from the author.

A.3.1 Discussion

The score developed above relies on a few assumptions:

- individual nucleotide matches can be considered independently of their surroundings;
- A, C, T, and G are equally likely to be conserved.

The parameter λ can be thought of as a kind of inverse distance measure: the closer it is to 0, the more randomized the sequences from which the matches were taken, and the closer it is to 1, the stronger the contribution from conservation. However, this parameter is essentially irrelevant: as long as it is held constant when comparing scores for different matches, the scores will be comparable, even for different background sequence distributions.

This score is easily extended for multiple sequence analysis, e.g. three-way comparison of matches. However, for each additional sequence, a new λ-like parameter is needed; it is unclear exactly what this means, however, in the context of actual scoring!

Finally, note that other than being readily calculable, the one major advantage this score has over the previously developed statistics for comparative sequence analysis is that it can easily take into account different background nucleotide probabilities, thus alleviating problems caused by e.g. AT-rich sequences in *C. elegans*.

Bibliography

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G., 1999. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press.